FUSION SCHEMES FOR ENSEMBLES OF HYPERSPECTRAL ANOMALY DETECTION ALGORITHMS

THESIS

Brooks R. Turnquist, Captain, USAF

AFIT-OR-MS-ENS-11-25

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT-OR-MS-ENS-11-25

FUSION SCHEMES FOR ENSEMBLES OF HYPERSPECTRAL ANOMALY DETECTION ALGORITHMS

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Brooks R. Turnquist, BS

Captain, USAF

March 2011

AFIT-OR-MS-ENS-11-25

FUSION SCHEMES FOR ENSEMBLES OF HYPERSPECTRAL ANOMALY DETECTION ALGORITHMS

Brooks R. Turnquist, BS
Captain, USAF

Approved:

//SIGNED// 21 MARCH 2011

Dr. Kenneth W. Bauer (Chairman) date

//SIGNED// 21 MARCH 2011

Dr. Mark A. Friend, Major, USAF (Member) date

AFIT-OR-MS-ENS-11-25

Abstract

Hyperspectral imaging is playing an ever increasing role in our military's remote sensing operations. The exponential increase in collection operations generates more data than can be evaluated by analysts unassisted. Anomaly detectors attempt to reduce this load on analysts by identifying potential target pixels which appear anomalous when compared to what are determined to be background, or non-target, pixels. However, there is no one individual algorithm that is best suited for all situations and it can be difficult to choose the best algorithm for each individual task. Fusion techniques have been shown to reduce errors and increase generalization, eliminating the need to always find the best algorithm for a given scenario. The utility of decision level fusion methods is examined, utilizing combinations of the emerging Autonomous Global Anomaly Detector and the Support Vector Data Description anomaly detection algorithms, along with the well-established Reed-Xiaoli detector. The fusion techniques investigated include algebraic combiners and voting methods. This research demonstrates that, with a modest amount of diversity among a minimal number of individual ensemble members, fusion offers reduced error rates and good generalization characteristics.

To my wife and two beautiful daughters.

Acknowledgments

I would like to express my sincerest gratitude to my thesis advisor, Dr. Kenneth Bauer, as well as my committee member, Major Mark Friend, for their guidance, support, and confidence in me throughout this endeavor.

In addition, I would be remiss if I did not acknowledge and thank Capt Jason Williams, Mr. Trevor Bihl, and my fellow classmates for all their support.

Brooks R. Turnquist

Table of Contents

List of Figures

List of Tables

FUSION SCHEMES FOR ENSEMBLES OF HYPERSPECTRAL ANOMALY DETECTION ALGORITHMS

I. Introduction

Hyperspectral imaging is playing an ever-increasing role in our military's remote sensing operations. The exponential increase in collection operations generates more data than can be evaluated by analysts unassisted. Anomaly detectors attempt to identify pixels which appear statistically anomalous when compared to what are determined to be background, or non-target, pixels. In doing this, anomaly detectors reduce the load on analysts by queuing them to these potential regions of interest. Many different algorithms have been proposed and are in use today, however, there is no one individual algorithm that is best suited for all situations and it can be difficult if not impossible to choose the best algorithm for each individual task. Often, an analyst is forced to simply choose an algorithm they are familiar with or have had luck with in the past, with no assurance that it is the right detector for the current situation. Fusion techniques have been shown to reduce errors and increase generalization by creating ensembles which capitalize on the diversity of the individual ensemble members. This reduces risk by eliminating the need to choose the best algorithm for a given scenario. The utility of decision level fusion methods is examined in the context of hyperspectral anomaly detectors, utilizing combinations of the emerging Autonomous Global Anomaly Detector and the Support Vector Data Description anomaly detection algorithms, along with the well-established Reed-Xiaoli detector. The fusion techniques investigated include algebraic combiners and voting methods.

II. Literature Review

HSI Basics

A digital camera captures information in the visible portion of the electromagnetic (EM) spectrum, usually in three bands or colors, red green and blue to form a digital photograph. That photograph, can be broken into smaller pixel elements or pixels based on spatial location. Hyperspectral Images are simply digital photographs which contain information from a much larger portion of the EM spectrum. This portion ranges from the ultraviolet to infrared wavelengths. (See Figure 1) Thus, instead of the three bands (red, green, and blue) that a normal color photograph has, a hyperspectral image can have tens to hundreds of contiguous bands. Each band in a hyperspectral image represents a small piece of the EM spectrum.

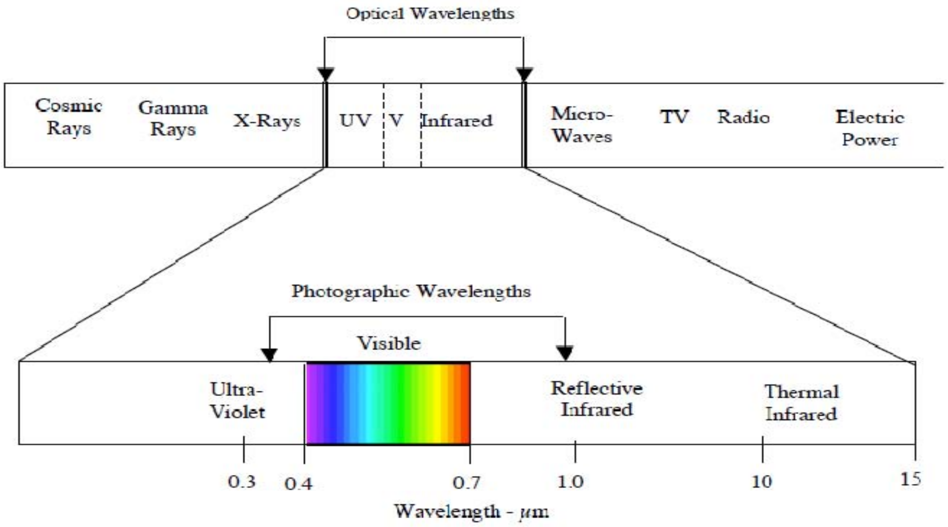


Figure 1: Electromagnetic Spectrum [1]

Data in a Hyperspectral image is stored in what is known as an image cube. An image cube is a three-dimensional data array where the m and n dimensions represent the spatial location (pixels) while the p dimension represents the spectral dimension (bands). The spectral dimension can be thought of as a stack of images, each capturing a small piece or band of the EM spectrum. The data representation described is depicted pictorially in Figure 2.

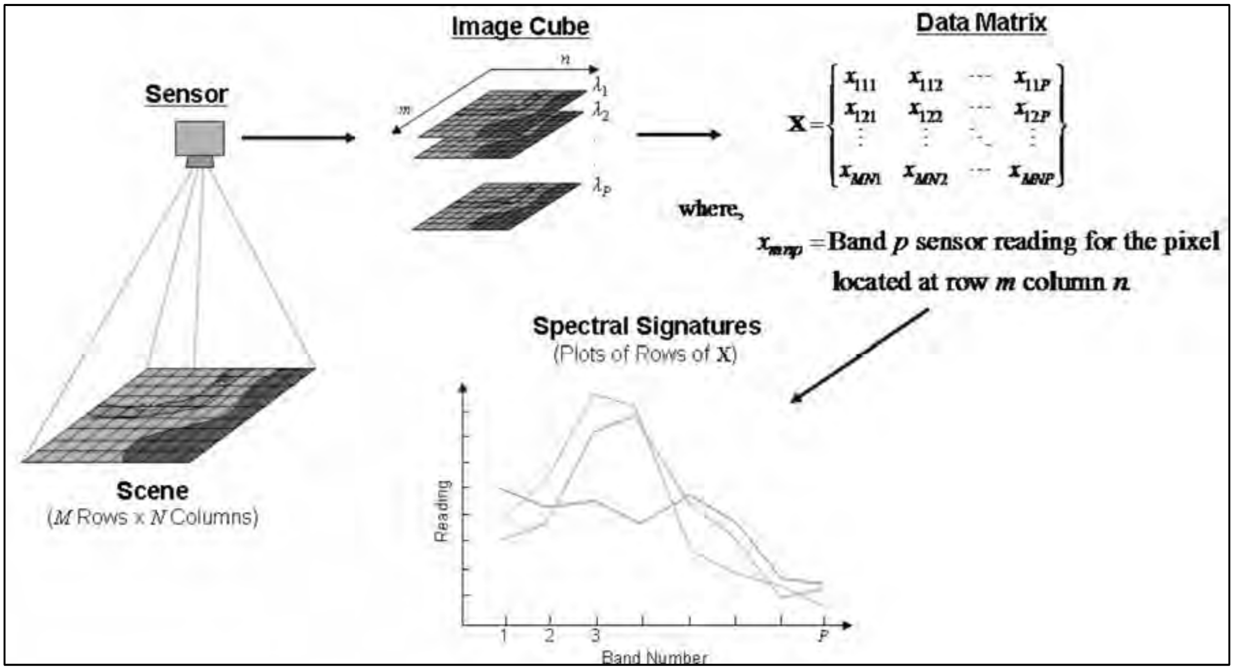


Image Cube
Data Matrix
Sensor
n
m
λ_1
λ_2
λ_p
$\mathbf{X} = \begin{Bmatrix} x_{111} & x_{112} & \cdots & x_{11P} \\ x_{121} & x_{122} & \cdots & x_{12P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{MN1} & x_{MN2} & \cdots & x_{MNP} \end{Bmatrix}$
where,
x_{mnp} = Band p sensor reading for the pixel located at row m column n
Spectral Signatures
(Plots of Rows of $\mathbf{X}$)
Reading
1
2
3
P
Band Number
Scene
(M Rows x N Columns)

Figure 2: The Basic Hyperspectral Imaging Process and Data Representation [2]

The basic task of hyperspectral imaging is to identify materials based upon their reflectance properties. [3] Since materials reflect light differently, this gives them their own unique spectral signature or fingerprint. This fingerprint is what allows us to differentiate a target of interest, say a tank, from the surrounding, background material, such as vegetation. This is why hyperspectral imaging is very useful in the field of remote sensing.

HSI Analysis Techniques

Spectral Matching techniques try to match a known spectral signature to pixels in the HSI. This can be quite difficult because the library of spectral signatures is usually given in reflectance units, while the sensor captures the radiance. The process of converting from radiance to reflectance or vice versa is complicated and requires some knowledge of the atmospheric conditions and viewing geometry at the time of the data collection [1] [2].

Anomaly Detection is an alternative to spectral matching and can be broken into two types; distribution based or global linear mixture models (LMM).[4] An anomaly is a pattern in the data which does not conform to a well-defined notion of what is normal. [5] The advantage to anomaly detectors is that they do not require prior knowledge of the target being searched for because they rely on the assumption of a target sparse environment and then simply search the data matrix for pixel vectors that are anomalous when compared to the remaining pixel vectors in the matrix using mean and covariance data. This eliminates the need for any radiance-reflectance conversions. On the other hand, an anomalous pixel vector flagged by an anomaly detector may not actually be anomalous but still must be analyzed further in order to identify the object. [2]

Distribution based methods can be further broken into local and global models. Local distribution based methods consider each pixel vector individually and compare its radiance values to the value of its surrounding pixels. The number of surrounding pixels can vary by detector but is generally from 8 to 50 of the nearest pixels. [4] Global mixture models are the other type of distribution based model and assume that each image contains some number, *K*, of different classes, or, as they are often referred to, endmembers. Each pixel is assumed to belong to one of those endmembers. The problem with and difficulty of this approach is determining the number of endmembers.

The second type of anomaly detection method is the global linear mixture model. This model, like the global distribution based model, assumes there are a certain number of endmembers, *C*, in the image. However, unlike the global distribution model, the global linear mixture model assumes that each observed pixel vector's spectrum is made up of a linear combination of the image's *C* endmember spectra. [3]

Reed-Xiaoli (RX) Detector

The RX algorithm is often thought of as the benchmark anomaly detector for hyperspectral imagery. It is a local anomaly detector and assumes Gaussian data which compares the Mahalanobis distance between a pixel under test and an estimate of the background mean to a threshold to detect an anomaly. [6] The test statistic the algorithm computes is given by:

$$RX(x) = (x-\hat{\mu})'\left(\frac{N}{N+1}S + \frac{1}{N+1}(x-\hat{\mu})(x-\hat{\mu})'\right)^{-1}(x-\hat{\mu})$$

where,

$\hat{\mu} = window\ mean\ vector$
$S = window\ covariance\ matrix$
$N = number\ of\ pixels\ in\ the\ processing\ window$

Using the Gaussian assumption to assess if a pixel is anomalous, the RX test statistic is compared to a threshold which is given by an appropriate quantile of the χ^2-distribution with *p* degrees of freedom for *p* dimensional data. [2]

There are a few known limitations to the RX detector's performance. Specifically, the local normal model assumption is inadequate in most situations and leads to poor false alarm performance. In addition, because it uses a local window approach, it has a hard time finding large anomalies. Finally, RX is computationally intensive when operating on hyperspectral imagery due to the need to estimate and invert large matrices.

Support Vector Data Description (SVDD)

The SVDD algorithm was originally proposed by [7] and its implementation in the area of hyperspectral anomaly detection was first proposed by [6]. The algorithm is a one-class support vector classifier that is able to estimate directly the support region for a data set. [6] utilizes SVDD to detect spectral anomalies that lie outside a region of support for a random vector. The SVDD algorithm does this by finding a minimum volume hypersphere about a set of random vectors, or in this case, the background pixels with each dimension corresponding to a different set of pixels. For a set of pixels,

$T = \{x_i: i = 1, \dots, M\}$ the algorithm seeks the smallest hypersphere in the induced feature space, $S = \{x: \|x - a\|^2 < R^2\}$ including the entire set T. This requires the following constrained optimization problem be solved:

$$\min(R) \;\; subject\ to\ x_i \,\epsilon\, S, i = 1, \dots, M$$

The corresponding Lagrangian with multipliers α_i can be optimized to find the center *a* and the radius *R*.

$$L(R, a, \alpha_i) = R^2 - \sum_i \alpha_i \left(R^2 - \langle x_i, x_i \rangle - 2\langle a, x_i \rangle + \langle a, a \rangle\right)$$

After optimizing *L* with respect to α_i and applying a kernel trick which employs the Gaussian radial basis function as the kernel function, the SVDD statistic simplifies to:

$$SVDD(y) = 1 - 2\sum_i \alpha_i K(y, \alpha_i) + \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) > \; R^2$$

$$\text{where, } \; K(x, y) = \exp\left(\frac{-\|x - y\|^2}{\sigma^2}\right)$$

The only free parameter, σ^2, in the radial basis function is the scale parameter and controls how well the SVDD generalizes to unseen data. [6][8]

The SVDD approach has the benefits of sparsity, requiring fewer training samples to characterize the background; being non-parametric, meaning it is data driven and avoids prior distributional assumptions about the data; and good generalization, avoiding over fitting and yielding good generalization results when compared to other classical methods. [6]

Autonomous Global Anomaly Detector (AutoGAD)

The AutoGAD is developed in [9] and is a four-phased algorithm used for detecting anomalies in hyperspectral imagery. The first two phases are feature extraction, the third stage is Feature Selection and the final stage is Identification.

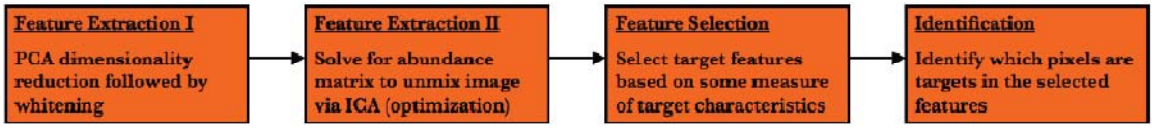

Feature Extraction I
PCA dimensionality reduction followed by whitening
Feature Extraction II
Solve for abundance matrix to unmix image via ICA (optimization)
Feature Selection
Select target features based on some measure of target characteristics
Identification
Identify which pixels are targets in the selected features

Figure 3: AutoGAD Process Flow for Target Detection [4]

The first feature extraction phase is a dimensionality reduction via principal components analysis (PCA). The dimensions are ordered according to the variance of the data captured by each component. The data is projected into this new principal component space where it is whitened. The number of dimensions to be retained is determined using the Maximum Distance Secant Line method proposed by [9].

The second feature extraction solves for the abundance matrix using independent component analysis. By projecting the data into independent components, [9] serves to comply with the Linear Mixture Model’s assumption that the components are independent.

Feature selection, takes the data and determines which features have potential to be targets by exploiting the assumption that targets are few and small in the scene. Using this, AutoGAD creates a potential target signal to noise ratio which is a ratio between potential targets and background pixels, which serves to set a noise floor.

Frequency histogram plots with signal to noise ratios above the noise floor represent potential target signatures.

The Identification phase looks at the plots with potential target signatures and chooses high scoring pixels as targets. AutoGAD mitigates its misclassification potential by employing a technique called Iterative Adaptive Noise (IAN) filtering. IAN filtering is an amplification technique which serves to give potential targets high contrast from the background, improving AutoGAD's designation capability. [4]

Fusion

The goal of fusion techniques is to extract complementary information from different sources to allow for a more informed decision than one could gain from any of the sources alone. [10] In theory, multisensor data fusion provides significant advantages over single sensor data. In addition to the statistical advantage, the use of multiple kinds of sensors may increase the possibility of a target of interest being observed and characterized resulting in a reduced error rate. In contrast, sensor fusion may not always result in an improved decision over simply selecting the most appropriate sensor for the task because accurate sensor data may be fused with very inaccurate data. [11]

Fusion Architectures

Three different architectures were proposed by the JDL Data Fusion Working Group, formed in 1986, based up on the actual level at which the information is fused together. [11] (See Figure 4) They are data level fusion, feature level fusion, and decision level fusion.

In data level fusion, raw data across the sensors are fused prior to any transformation or identity declarations. After fusing the data, the identification process proceeds like it would for a single sensor. Data level fusion is illustrated in Figure 4(c) and is often referred to as pixel level fusion when dealing with imagery data. [12] This architecture can only be directly used when the sensor data being fused are from identical or commensurate sensors. [11][12] The extent to which this data-level fusion architecture can be employed with diverse data types depends on the availability of an accurate physical model. For example, synthetic aperture radar images can be fused with visible images if the appropriate corrections can be modeled, accounting for viewing geometry, scaling, and other related factors.[12]

Feature level fusion is depicted in Figure 4(b). In this architecture, features are combined into a single, joint feature vector which then serves as an input to a classification technique. Prior to concatenation, image registration and data association must be performed to ensure that the information being fused relates to the same object. [11] Techniques available to perform the identification process include cluster analysis, neural networks, and knowledge based techniques. Feature vectors must be sorted into meaningful groups in this approach using some type of association process since the vectors may be vastly different quantities. [12]

In the decision level fusion approach, shown in Figure 4(a), each sensor collects its own measurement and transforms it into a decision space variable regarding a presence or absence of targets. These decision space variables are then combined from the multiple sensors to make a joint identity declaration regarding the presence or absence of targets. [11] Techniques used for fusing decision space variables include voting methods, algebraic combiners, Bayesian inference, Dempster-Shaffer's method, generalized evidence processing theory, and various ad hoc methods. [10][11][12]

In general, better accuracy is obtained the closer the fusion is performed to the source. This means that data level fusion is likely the most accurate, and then feature level fusion, then decision level fusion. However, data level fusion is limited to sensors collecting the same types of data and requires an accurate data alignment among the sensors. Fusion at the decision level does not require identical or commensurate sensors because it involves fusing the individual classifier outputs. Other factors that must be taken into account when choosing a fusion approach involve the implications for system implementation. Thus, the choice of which fusion approach to use must be a system level decision. [12]

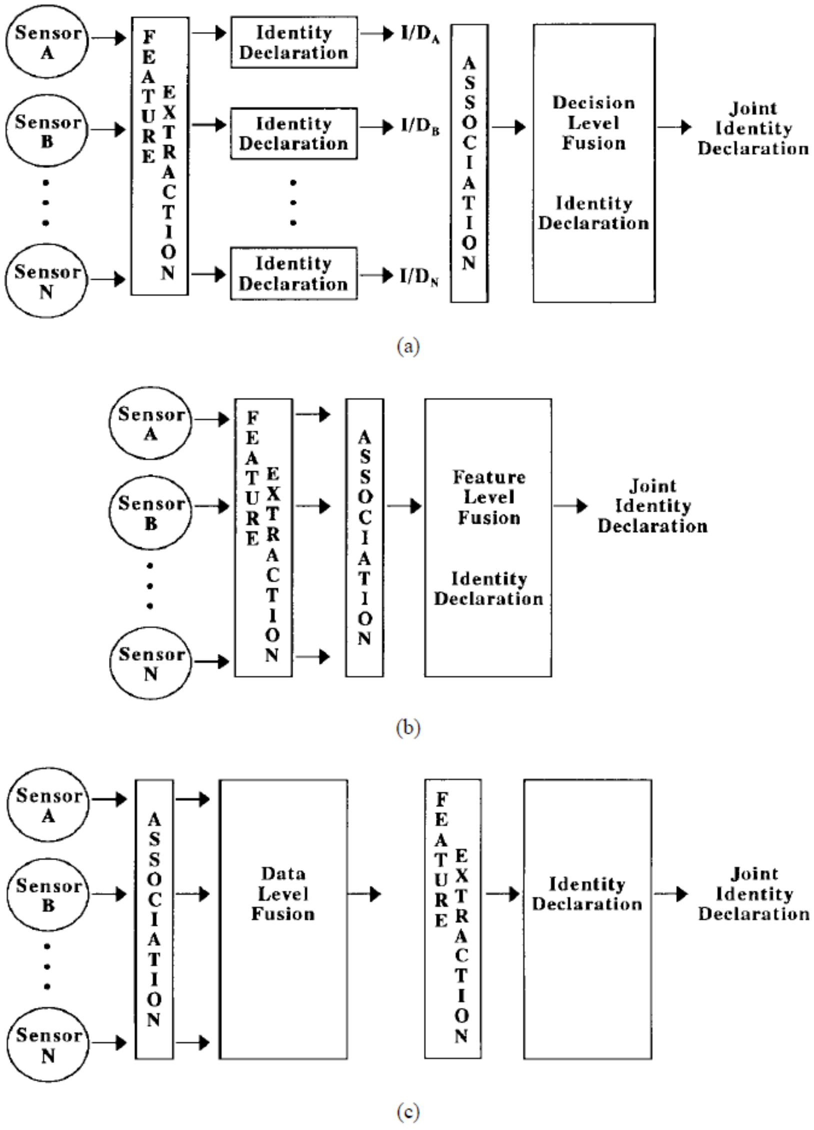


Figure 4: Alternate Architectures for Multi-sensor Data Fusion [11]

Decision Level Fusion Schemes

There are a myriad of decision level fusion schemes in the literature but most can be categorized as one of two types: Those that combine class labels and those that combine the outputs of continuous classifiers. Voting methods, which combine class labels, and Algebraic combiners, which combine outputs of continuous classifiers, are the two most common methods found in the literature. Other methods described in the following paragraphs include Behavior Knowledge Space, Borda Counts, Bayesian Inference, and Dempster-Shaffer Theory.

Voting Methods

Voting methods are generally appropriate for combinations of classifiers whose outputs are categorical for a classification problem. [13] These methods use a democratic process and treat each individual sensor's identity declaration as a vote, combining them to get a joint identity declaration. [12] There are a number of voting methods that can be used to combine classifiers. Three of the most commonly used methods are 1) unanimous voting where all classifiers must agree; 2) simple majority voting where the winning class must be predicted by at least one more than half of the classifiers; and 3) plurality or majority voting where classification is determined by the class that receives the highest number of votes, regardless of if it gets half.[10] For plurality voting among T classifiers and C classes, the decision can be modeled as:

$$\begin{aligned} &choose\ class\ \omega_J\ if: \\ &\qquad \textstyle\sum_{t=1}^{T} d_{t,J} = \max_{1\le \mathrm{j} \le \mathrm{C}} \sum_{t=1}^{T} d_{t,j} \\ &where, \\ & \\ &d_{t,j} = decision\ of\ t^{th}\ classifier\ \epsilon\ \{0,1\} \\ &\quad t = 1, \dots, T \\ &\quad j = 1, \dots, C \end{aligned}$$

It can be shown that voting is an optimal combination scheme under a few minor assumptions: 1) an odd number of classifiers exist for a two-class problem; 2) the probability, *p*, of each classifier choosing the correct class for any instance *x* is greater than 0.5; and 3) classifier outputs are independent. [10] In addition, weights can be employed to these methods to try to account for differences in sensor performance. These weights can be computed using signal-to-noise ratios and other factors. [12]

Behavior Knowledge Space

The Behavior Knowledge Space (BKS) is another method that can be used to combine class labels. The BKS was originally proposed by Huang and Suen and uses look up tables constructed from training data classifications to keep track of the frequency of which each labeling combination is used. [10] Figure 5 illustrates how the BKS works. This combination scheme looks at all possible combinations of the labeling for classifiers and picks the most observed true class for each combination of labels. In this example, since the combination $\omega_1, \omega_2, \omega_1$ occurs in the training data for the true class of ω_2 the most number of times, when this combination is seen in the future, ω_2 will be determined the wining class.

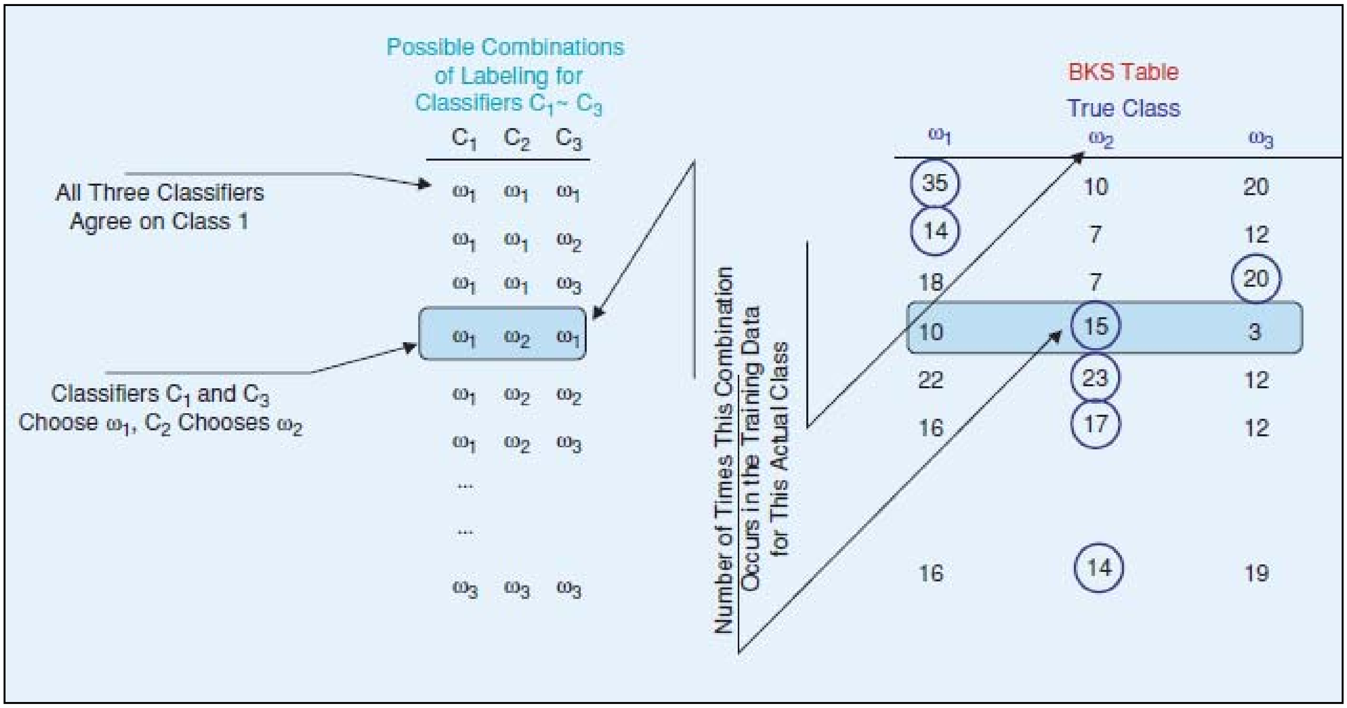

Possible Combinations of Labeling for Classifiers C_1~ C_3
C_1 C_2 C_3
All Three Classifiers Agree on Class 1
ω_1 ω_1 ω_1
ω_1 ω_1 ω_2
ω_1 ω_1 ω_3
ω_1 ω_2 ω_1
Classifiers C_1 and C_3 Choose ω_1, C_2 Chooses ω_2
ω_1 ω_2 ω_2
ω_1 ω_2 ω_3
...
...
ω_3 ω_3 ω_3
Number of Times This Combination Occurs in the Training Data for This Actual Class
BKS Table
True Class
ω_1 ω_2 ω_3
35 10 20
14 7 12
18 7 20
10 15 3
22 23 12
16 17 12
16 14 19

Figure 5: Behavior Knowledge Space (BKS) Illustration [10]

Borda Count

The Borda Count was developed in 1770 by Jean Charles de Borda and is typically used when classifiers have the ability to rank order the classes. It is different from previously discussed methods in that it does not discard the support of the non-chosen classes. [10] The Borda Count works as follows for each classifier: the top ranked class (out of C classes) receives *C-1* votes, the second ranked gets *C-2* votes, on down to the lowest ranking class receiving zero. The class that receives the most votes across all the classifiers is deemed the winner.

Algebraic Combiners

Algebraic combination schemes suit systems where the classifier outputs are continuous in nature. The continuous outputs provided by a classifier are interpreted as the support given to that class, and is generally accepted as an estimate of the posterior probability for that class. Common algebraic combination schemes are the mean, trimmed mean, min/max/median, and product rules. In each of these methods, the total support for each class is obtained by a simple function of the support received by each individual classifier. [10]

Using the average or mean rule, the total support for each class is obtained as the average of all T classifiers' *jth* outputs, and can be described as:

$$\mu_j(\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T} d_{t,j}(\boldsymbol{x})$$

$$where,$$

$$d_{t,j} = support\ of\ t^{th}\ classifier\ for\ j^{th}\ class\ \epsilon\ \{0,1\}$$
$$t = 1, \dots, T$$

The decision is made as the class ω_j for which the support μ_j is the largest. [10] [13][14]

Much like the voting methods discussed earlier, a weighting system can be applied to the average rule, but in this case, the weights are not applied to the class label but to the actual continuous outputs. Weightings can be obtained in a trainable or non-trainable manner. [10]

Another algebraic combiner is the trimmed average. This method removes the most optimistic and pessimistic classifiers before calculating the average. This method removes the adverse affects of classifiers that give unusually high or low support to a given class. [10]

The Minimum/Maximum/Median Rule simply takes the minimum, maximum, or median among all the classifiers' outputs and the decision is made as the class for which the most support is given. [10]

The product rule multiplies the supports given by the classifiers. This method is extremely sensitive to supports close to zero, effectively removing any chance of that class being chosen. Nevertheless, if the posterior probabilities are estimated accurately, the product rule will provide the best estimate of the overall posterior probability of the class selected. [10]

$$\mu_j(\boldsymbol{x}) = \frac{1}{T}\prod_{t=1}^{T} d_{t,j}(\boldsymbol{x})$$

$where,$

$$d_{t,j}(x) = support\ of\ t^{th}\ classifier\ for\ j^{th}\ class\ \epsilon\ \{0,1\}$$
$$t = 1, \dots, T$$

Bayesian Inference

Bayesian inference is a modification to classical inference techniques and is more commonly used in identity fusion. [12] Bayesian inference methods address some of the difficulties with the classical techniques by using prior likelihoods and newly observed evidence information to obtain updated posterior likelihoods. Bayes uses the classifier identity declarations and the *a priori* data to give a conditional probability for each class type given the declarations from the sensors. By applying the maximum *a posterior* probability rule, the conditional probability that has the largest value is chosen.

Advantages of Bayes' technique are that it determines the probability of a declaration being correct given the evidence. It also allows for incorporation of *a priori* knowledge about the likelihood of an individual declaration being correct in the absence of data, and finally, it has the ability to use subjective probabilities for the *a priori* probabilities of classifier declarations and probabilities of evidence given class. [12]

Some of the disadvantages of Bayes' fusion technique are that the prior likelihood functions are difficult to define, it can become complex when multiple classes and multiple dependent events exist, it requires that classes be mutually exclusive, and there is no measure of uncertainty. [12]

Dempster-Shafer Theory

Dempster-Shafer theory is a generalization of Bayesian Inference methods and utilizes belief functions to quantify evidence from each source, instead of probabilities, which are combined using Dempster's rule of combination. [10] Whereas Bayesian theory requires probabilities for each class of interest, belief functions allow the user to base the degree of belief of some question based on the probability of a related question. [15] Dempster-Shafer theory does not require that mutually exclusive classes be defined, but rather propositions which may contain overlapping or conflicting classes. [12] The theory of belief functions is based on two ideas: (1) a degree of belief about one question can be obtained from subjective probabilities of a related question and (2) combining the degrees of belief using Dempster's rule of combination when based on independent evidence. [15] Shafer offers up the following example in [15] to help illustrate this concept. Suppose you have subjective probabilities for your friend

Betty's reliability. The probability that she is reliable is 0.9 and the probability that she is unreliable is 0.1. Now suppose that she tells you that a tree fell on your car. If she is reliable, this statement must be true. However, if she is unreliable, this statement is not necessarily false. Her statement justifies a 0.9 belief that a tree fell on your car, but a zero degree of belief that no tree fell on your car. A zero degree of belief does not mean that you are sure no tree fell on your car; it means that Betty's statement does not give you any reason to believe that no tree fell on your car. These two beliefs of 0.9 and zero together constitute a belief function.

Dempster's rule for combination is based on probabilistic independence. Continuing with Shafer's example, say you have a second friend, Sally, who also says that a tree fell on your car. Since Betty's reliability is independent of Sally's, then you can simply multiply the probabilities of events. Assuming that your belief in Sally's reliability is the same as Betty's, this result gives the probability that both are reliable to be 0.9 x 0.9 = 0.81, the probability that neither is reliable to be 0.1 x 0.1 = 0.01, and the probability that at least one is reliable to be 1 – 0.01 = 0.99. However, if Sally and Betty make contradictory statements (Betty says a tree fell on your car and Sally says no tree fell on your car), then some of the joint probabilities computed for who is reliable must be ruled out and the remaining probabilities normalized to sum to one. Once these probabilities are normalized, new degrees of beliefs can be calculated.

Shafer notes that the net result of Dempster combination is that evidence that agrees reinforces, contradictory evidence erodes, and a chain of evidence is weaker than its weakest link. [15]

The renormalization required in light of contradictory evidence is called probabilistic conditioning and can destroy the initial assumption of independence. This is one of the biggest criticisms of the Dempster-Shafer theory. [12]

Other classifier combination methods found in the literature, which are not described in this thesis, include Decision Templates, Singular Value Decomposition, Wernecke's Method, First-Order Dependence Trees, and Fuzzy Integrals. There is no unique combiner that is best suited for all problems. However, the weighted average rule has been the most widely used due to its consistently good performance and ease of implementation. In some cases, the simple average has rivaled the weighted average's performance. [16]

Diversity

The performance of a model using a fusion scheme can be influenced by a number of factors involved in the construction and operation of the model. These factors consist of the number of classifiers used, the accuracy of the individual classifiers, and the diversity among the classifiers. [10][13]

Diversity is an estimate of the difference of making the same errors between models. [13] A set of classifiers is said to be diverse if the classifiers' decision boundaries are adequately different from those of others. [10] This means that each classifier is as unique as possible, especially with respect to which observations are misclassified. We would like our classifiers to be as correct as possible, but in the case that there are errors, we would like them to be on different instances. There are many diversity measures but they can be categorized as being either pair-wise or non-pair-wise. Pair-wise measures are the simplest and estimate the diversity between two classifiers while the non-pair-wise measures estimate the diversity of all the classifiers coincidentally. [13] For the pair-wise diversity measures, the following notations describe the relationship between two classifiers, *i* and *j*,

Table 1: Notation for pair-wise relationships between two classifiers

	Classifier j is correct	Classifier j is incorrect
Classifier i is correct	a	b
Classifier i is incorrect	c	d

where a is the fraction of instances that are correctly classified by both classifiers, b is the fraction correctly classified by only the *ith* classifier, *c* is the fraction classified correctly by only the *jth* classifier, and d is the fraction of instances that were incorrectly classified by both classifiers.

Correlation diversity measures the correlation between two binary classifier outputs and is defined as

$$\rho_{ij} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

When ρ=0, indicating uncorrelated classifiers, maximum diversity is obtained. [10][16][17][18]

Yule's Q-statistic is probably the most simple metric and is defined as

$$Q_{ij} = \frac{ad - bc}{ad + bc}$$

If the classifiers tend to make the same classification of observations the Q-statistic assumes positive values, otherwise it assumes negative values. Again, when Q=0, maximum diversity is obtained. For any two classifiers, Q and ρ have the same sign, and it can be proven that $|\rho| \leq |Q|$. [10][16][17][18]

Disagreement measures the amount that the two classifiers make differing classifications while the Double Fault measures the amount of time both classifiers are incorrect. Diversity increases with increases in Disagreement and decreases in the Double Fault Measure. [10] [16] [17][18]

$$D_{ij} = b + c$$
$$DF_{ij} = d$$

The entropy measure assumes that diversity is highest when half of all the classifiers are correct and the remaining half are incorrect. The measure of entropy is defined as

$$E = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{T - \left\lceil \frac{T}{2} \right\rceil} \min \{\zeta_i, (T - \zeta_i)\}$$

where ζ_i is the number of classifiers, out of *T*, which misclassifies an observation and *N* is the number of observations. An entropy value of zero means that there is no diversity and a value of one means there is maximum diversity. [10][17][18]

The Kohavi-Wolpert variance is similar to the disagreement measure and can be defined as

$$KW = \frac{1}{NT^2}\sum_{i=1}^{N} \zeta_i \cdot (T - \zeta_i)$$

Diversity is said to increase as the Kohavi-Wolpert variance increases. It can also be shown that the Kohavi-Wolpert variance differs from the averaged disagreement measure by a factor of $\frac{T-1}{2T}$. [10][17][18]

The measure of difficulty, θ, is the measure of variance of a random variable $Z(x_t) = \{0,1/T, 2/T, \ldots ,1\}$ where Z_t is the fraction of classifiers that misclassify an observation x_t and $\bar{z}$ is the mean of *Z.*

$$\theta = \frac{1}{T}\sum_{t=0}^{T}(z_t - \bar{z})^2$$

This measure was originally proposed by [19] who argued that if classifiers tend to correctly classify and misclassify the same instances, θ becomes large and there is little diversity. If the classifiers misclassify different instances, then the measure is small and the diversity is high. So, as the measure of difficulty increases, so does the diversity. [10][17][18][19]

Consider a two-class problem where three classifiers have made identity declarations of ten observations. Error! Reference source not found. **shows the declarations made by each classifier as well as the true class for each observation. Table 3,**

Table 4, and

Table 5 show the pair-wise relationships between the three classifiers.

Table 6 shows the resulting measures of diversity for this three-classifier fusion problem.

Table 2: Identity declarations for a two class, 3 classifier problem over 10 observations

Observation Number	Classifier 1 Identity Declaration	Classifier 2 Identity Declaration	Classifier 3 Identity Declaration	True Class
1	0	1	1	1
2	0	0	0	1
3	0	1	0	0
4	1	1	1	0
5	1	1	1	0
6	0	0	1	0
7	1	1	1	1
8	0	0	0	1
9	0	1	1	1
10	0	0	0	0

Table 3: Pair-wise relationship between classifiers 1 and 2

	Classifier 2 Correct	Classifier 2 Incorrect
Classifier 1 Correct	0.30	0.10
Classifier 1 Incorrect	0.20	0.40

Table 4: Pair-wise relationship between classifiers 1 and 3

	Classifier 3 Correct	Classifier 3 Incorrect
Classifier 1 Correct	0.30	0.10
Classifier 1 Incorrect	0.20	0.40

Table 5: Pair-wise relationship between classifiers 2 and 3

	Classifier 3 Correct	Classifier 3 Incorrect
Classifier 2 Correct	0.40	0.10
Classifier 2 Incorrect	0.10	0.40

Table 6: Comparison of Measures of Diversity

	Measures of Diversity						
	Correlation Diversity	Yule's Q Statistic	Disagreement	Double Fault	Entropy	Kohavi-Wolpert	Measure of Difficulty
1, 2	0.43	0.71	0.30	0.40	-	-	-
1, 3	0.43	0.71	0.30	0.40	-	-	-
2, 3	0.60	0.88	0.20	0.40	-	-	-
Overall	0.45	0.60	0.26	0.40	0.40	0.09	0.53

Other measures of diversity found in the literature include interrater agreement, generalized diversity, and coincident failure diversity and are described in [13] [16] [17] and [18]. There are many different definitions of and metrics for diversity, and while there is no clear best method, there does seem to be some consensus that Yule's Q-statistic is the best choice when no additional information is available due to its ease of interpretation and implementation. [10][17]

III. Methodology

Method

This thesis investigates the fusion of all two-member ensembles of AutoGAD, SVDD, and RX detector outputs as well as the three member ensemble using the continuous and voting fusion methods listed below to determine potential benefits.

- Continuous Methods
    - Maximum
    - Mean
    - Product
- Voting Methods
    - Unanimous Voting
    - Majority Voting (requires at least three inputs)

The methods that require training, such as the weighted average, were not implemented here due to a lack of sufficient data to train and test them adequately.

AutoGAD has a number of user specified settings that can be changed and manipulated. The settings used for this analysis are those that were originally proposed in [9] and are shown in

Table 7.

Table 7: AutoGAD Settings Used

funct=2;	objective function in ICA to use. Options [1=tanh, 2=pow3]
orthog=1;	find ICs in parallel (symm) or one by one (defl).
	Options [symm=1, defl=2]
dim_adjustment=0;	how much to adjust max distance log scale secant line (MDLS)
	dimensionality decision
max_score_thresh=10;	threshold above which decision is made to declare target
bin_width_SNR=.05;	bin width when using zero-detection histogram method to
	determine breakpoint between background and potential targets for
	calculating potential target SNR (PT SNR)
bin_width_ident=.05;	bin width when using zero-detection histogram method to
	determine breakpoint between background and targets for identifying target
	pixels from selected target signals
threshold_both_sides=0;	1=identifiy outliers on both sides of IC signal,
	0=identify ouliers on side with highest magnitude scores only
clean_sig=1;	0 = no signal smoothing, 1 = signal smoothing prior to target
	identification
smooth_iter_high=10;	number of iterations to complete for iterative smoothing
	of low SNR object
smooth_iter_low=20;	number of iterations to complete for iterative smoothing
	of high SNR object
low_SNR=10;	Threshold decision for choosing smooth_iter_low or smooth_iter_high
window_size=3;	image window size for smoothing
iteration_coeff = 50;	
PT_SNR_thresh=2;	threshold above which decision is made to declare target
req_corr = 0.98514236;	Threshold correlation required for bands to be clustered together
Kurtosis_thresh=9;	threshold above which decision is made to declare target
target_fraction_thresh = 0.0269;	The maximum fraction of the image expected to contain target pixels.
Left_Kurt_Thresh=9;	If left side kurtosis is less than threshold program will
	not perform thresholding on both sides for that map

SVDD has two parameters that can be specified by the user, the scale parameter σ^2, which sets the tightness of fit, and the number of random pixels N to be used as an estimate of the background. For this analysis, the scale parameter is set to the minimax estimate of 905 suggested by the work of [8], with a sample size of for the background spectra of 500 pixels. For the RX algorithm, user specified parameters are window size which was set to 25x25 pixels and the number of dimensions to keep which was set to seven. It must be acknowledged that, in using these near-optimal settings for the individual algorithms, the settings are not being allowed to interact in a way that could allow the fusion to perform at potentially even higher levels.

The idea with fusion is to create ensembles of anomaly detectors with a high degree of diversity to maximize the benefit from the fusion scheme. The fusion of these three anomaly detectors is investigated because they each have different approaches to anomaly detection and therefore, hopefully, each detector would provide useful pieces of information that the others could not. AutoGAD and SVDD are global methods while RX uses a local window approach. In addition, AutoGAD and SVDD are emerging anomaly detectors while RX has been a benchmark in the field.

To get a measure of the diversity among the AutoGAD, SVDD, and RX algorithms, Yule's Q-statistic, Correlation, Disagreement and Double Fault pair-wise diversity measures were calculated as well as an overall diversity measure for the three member ensemble which utilizes the identity declarations made by each algorithm when they were falsely identifying roughly 10 percent of the pixels as anomalies. Figure 6 demonstrates the pair-wise comparisons generated for the ARES1F image. Using this information, the pair-wise measures of diversity are calculated and provided in Table 8. The overall measures of diversity listed are simply an average of the three pair-wise diversity measures.

	RX Correct	RX Incorrect
AutoGAD Correct	0.8020	0.1079
AutoGAD Incorrect	0.0820	0.0081

	SVDD Correct	SVDD Incorrect
AutoGAD Correct	0.8365	0.0734
AutoGAD Incorrect	0.0710	0.0191

	SVDD Correct	SVDD Incorrect
RX Correct	0.8183	0.0657
RX Incorrect	0.0893	0.0267

Figure 6: AutoGAD, SVDD, and RX pair-wise relationships for ARES1F when FPF = 10%

Table 8: Pair-wise measures of diversity between AutoGAD, RX, and SVDD for ARES1F when FPF = 10%

	Yule's Q-statistic	*Correlation*	*Disagreement*	*Double Fault*
Pair-wise between AutoGAD and RX	-0.1502	-0.0251	0.1898	0.0081
Pair-wise between AutoGAD and SVDD	0.5072	0.1294	0.1444	0.0190
Pair-wise between RX and SVDD	0.5770	0.1725	0.1550	0.0267
Overall Measure	0.3113	0.0922	0.1630	0.0179

Maximum diversity when using Yule's Q-statistic and Correlation is obtained at a value of zero. When using the disagreement measure a higher value indicates a higher diversity while a lower double fault measures indicates a greater amount of diversity. As can be seen in Table 8, this combination of classifiers exhibits a much lower amount of diversity than would be preferred according to Yule's Q-statistic and the disagreement measures, while the correlation and double fault measures seem to suggest a higher level of diversity. Figure 7 compares the anomaly declaration masks for each method as compared to the truth. Here it can be seen that both AutoGAD and SVDD seem to be picking up and identifying the same observations as anomalies, which is leading the pair-wise diversity measure between the two to indicate a lack of diversity. RX, however seems to be misclassifying observations much differently than

AutoGAD, resulting in pair-wise diversity measures which lean towards a higher amount of diversity. While these levels of diversity exhibited by the combinations of AutoGAD, SVDD, and RX may not be as high as would be preferred, the benefits of fusion will be explored realizing that the effects may not be as significant as they would be if a more diverse set of anomaly detectors had been chosen.

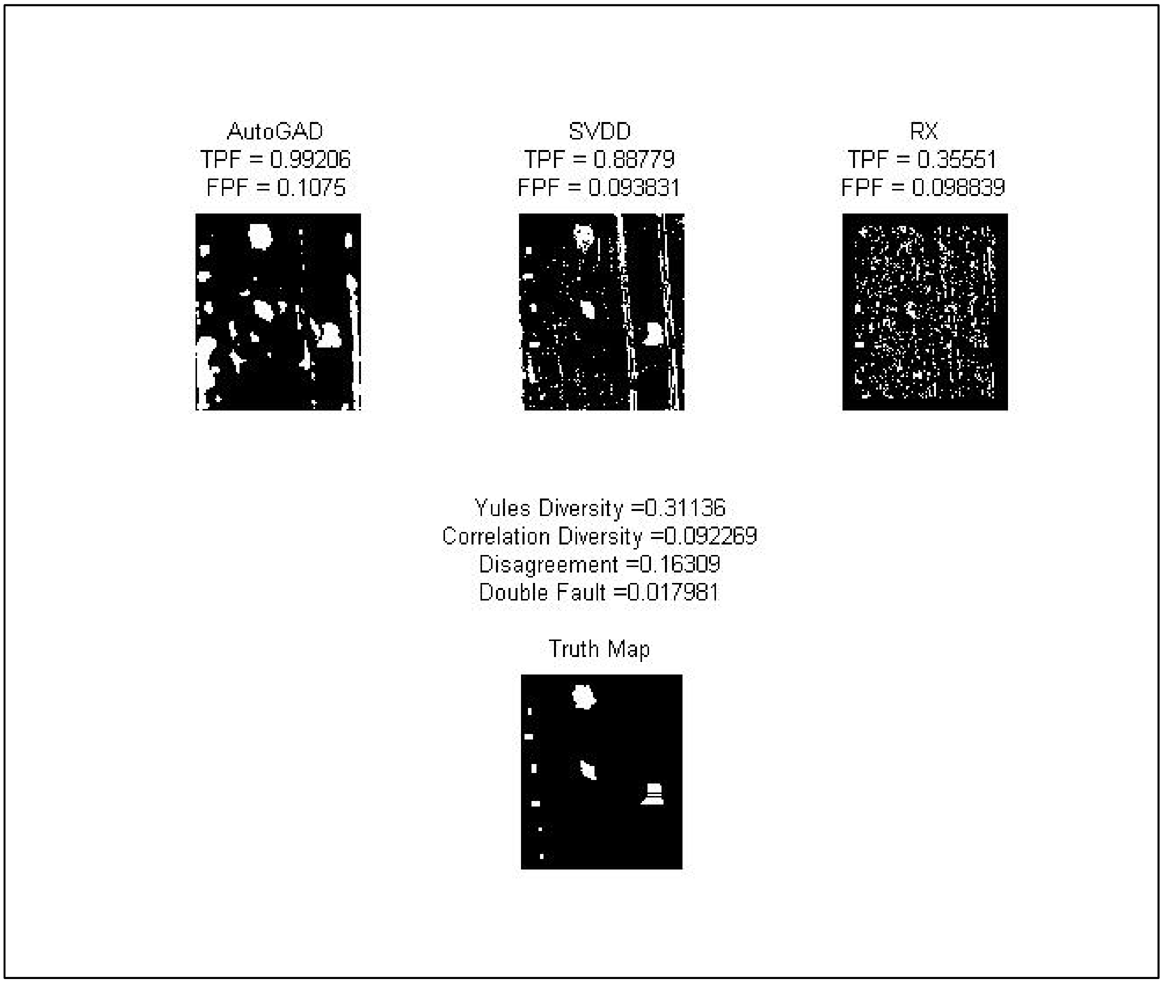


Figure 7: Masks for AutoGAD, SVDD, RX and Truth for ARES1F with calculated diversity measures

Now that the diversity among the anomaly detectors has been calculated and the combination of classifiers chosen, the outputs can be fused. Normally, when using the continuous methods, the inputs to the fusion scheme need to be some measure of support for each class. However, the three methods investigated here do not readily provide scores which can be regarded or interpreted as the support for each class. Moreover, while the SVDD and RX algorithms provide a single statistic for each pixel that could be interpreted as the support for a pixel being anomalous, AutoGAD does not. AutoGAD actually provides multiple scores for each pixel, the number of scores being equal to the number of independent components which highlight potential targets. Furthermore, these scores are not commensurate in magnitude. In order to obtain a single score from AutoGAD for each pixel, the following two step procedure is used:

1) Normalize pixel scores for each feature by dividing by the respective identity threshold. This makes all potential target pixels in a feature greater than one and background pixels less than one.
2) Retain only the maximum score for each pixel across all normalized feature vectors.

Figure 8 illustrates how this procedure works on a simple two-dimensional example with only five pixels.

$$Original\ AutoGAD\ scores = \begin{bmatrix} 4 & 10 \\ 3 & 12 \\ 5 & 11 \\ 1 & 15 \\ 2 & 9 \end{bmatrix} \quad AutoGAD\ Target\ Thresholds = [4\ 13]$$

$$Normalized\ AutoGAD\ Scores = \begin{bmatrix} 1.00 & 0.77 \\ 0.75 & 0.92 \\ 1.25 & 0.85 \\ 0.25 & 1.15 \\ 0.50 & 0.69 \end{bmatrix} \quad \xrightarrow{Max\ Score} \quad AutoGAD\ Single\ Pixel\ Score = \begin{bmatrix} 1.00 \\ 0.92 \\ 1.25 \\ 1.15 \\ 0.69 \end{bmatrix}$$

Figure 8: Example illustrating AutoGAD single pixel score procedure

Once a single pixel score has been attained for each of the three methods, they are normalized onto a zero to one scale using the formula below to make each method's continuous outputs proportionate.

$$\tilde{x}_{\boldsymbol{i}} = \frac{x_i - \min(x)}{\max(x) - \min\,(x)};for\ i = 1, \ldots, Num\ Pixels$$

Now all the detectors have scores that are on the same scale which can be interpreted as the support of that detector for that pixel being anomalous; but there is still no score which can be interpreted as that detector's support of a pixel being part of the background. Normally these fusion schemes use the comparison of these supports to choose the winning class. In the absence of this additional score, the support scores are fused together using the mean, max, or product of the scores and then simply compared to a decision threshold which is somewhere on the interval of zero to one to identify anomalous pixels.

Measures

The following measures are what will be used in comparing the average performance of each method. The True Positive Fraction (TPF) is calculated as the number of correctly identified target pixels divided by the total number of target pixels. False Positive Fraction (FPF) is calculated as the number of incorrectly identified non-target pixels divided by the total number of non-target pixels. It should be noted, in relative TPF and FPF calculations, the pixels bordering targets were counted as True Positives if AutoGAD classified them as a target pixel and as a True Negative if AutoGAD classified them as background pixels. This method of calculating the TPF and FPF makes the ROC curve move up and to the left, inflating the apparent performance of the algorithm. In this thesis, these border pixels were simply ignored for the TPF and FPF calculations.

In order to compare the trade-offs between TPF and FPF for each anomaly detector, receiver operating characteristic (ROC) curves are used which plot TPF versus FPF. Developing these curves for the SVDD and RX algorithms is straightforward in that we can simply vary the threshold by ranging the alpha level from 0 to 1. However, AutoGAD calculates its own thresholds using a variety of parameters and thus creating a ROC curve is not so straightforward. In order to gain a ROC curve on the performance of AutoGAD, the target identity threshold was varied as a percentage of the original threshold calculated by the AutoGAD algorithm. However, an equivalent ROC curve is generated by using the new AutoGAD score described earlier and varying the single decision threshold. Figure 9 shows the equivalence between the two methods. It should also be noted there are likely many different values which could be varied in AutoGAD to obtain a ROC curve, however this is the one, which will be used here.

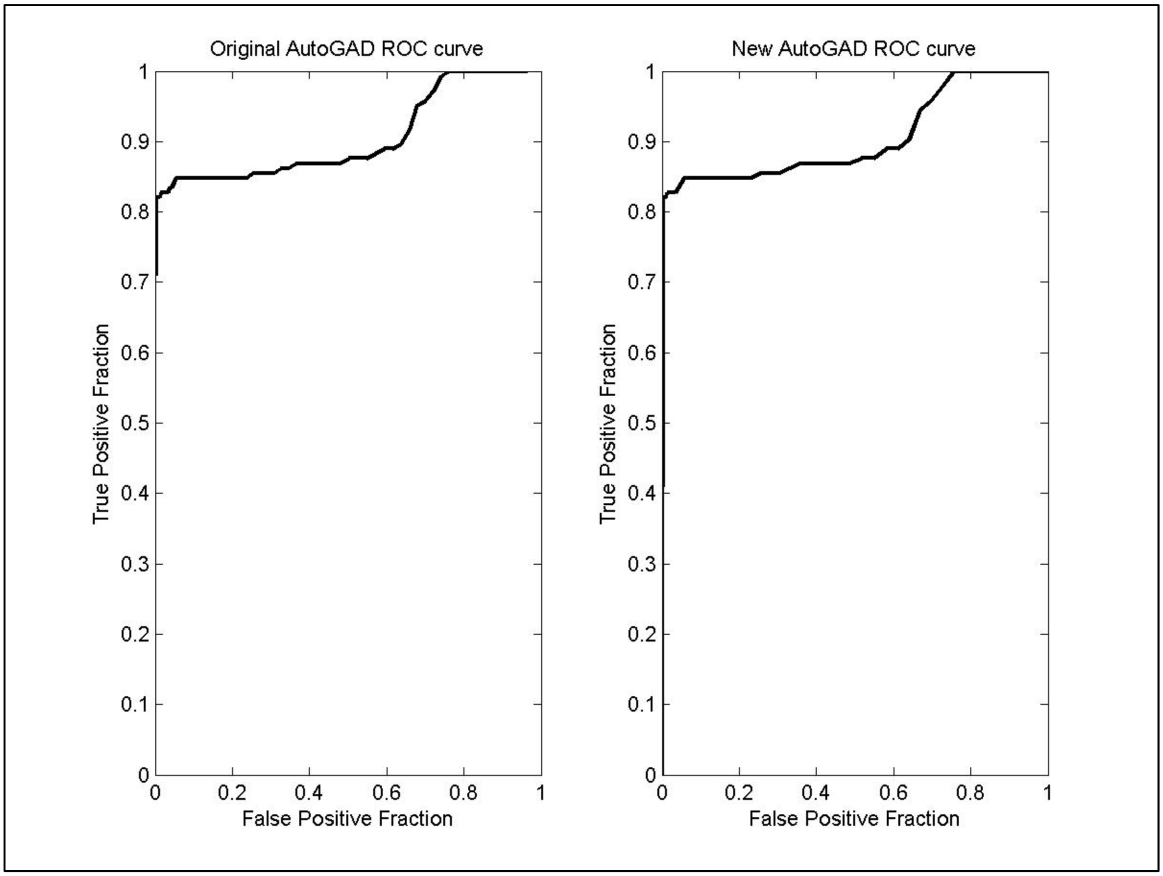


Figure 9: Comparison of ROC curves to show equivalence between the original AutoGAD score and the new AutoGAD score on ARES3F image

To combine the class outputs of each individual detector using the voting methods, thresholds have to be applied prior to the fusion rule being employed. This research varies AutoGAD's threshold from 0.1 to 4.5 times its internally calculated threshold, in increments of 0.01. RX and SVDD's alpha values vary from zero to one in increments of 0.01. The exploration of all the possible combinations of thresholds when analyzing a single image yields a family of ROC curves for each voting method. Figure 10 shows how a family of ROC curves might look for a single voting method on any given image. Each point in the family of curves corresponds to a unique combination of threshold settings. The black points make up what is referred to as the ROC manifold and represent the pareto-optimal setting combinations. The points which fall down and to the right of the manifold are inferior threshold setting combinations because other settings exist which provide improved performance. These inferior points are removed from the ROC curve and only the manifold is used for comparison purposes. Families of curves are not generated when applying the continuous fusion rules because the threshold is not applied until after the individual support scores are fused. Therefore, only one threshold need be applied in these cases and only one ROC curve is generated.

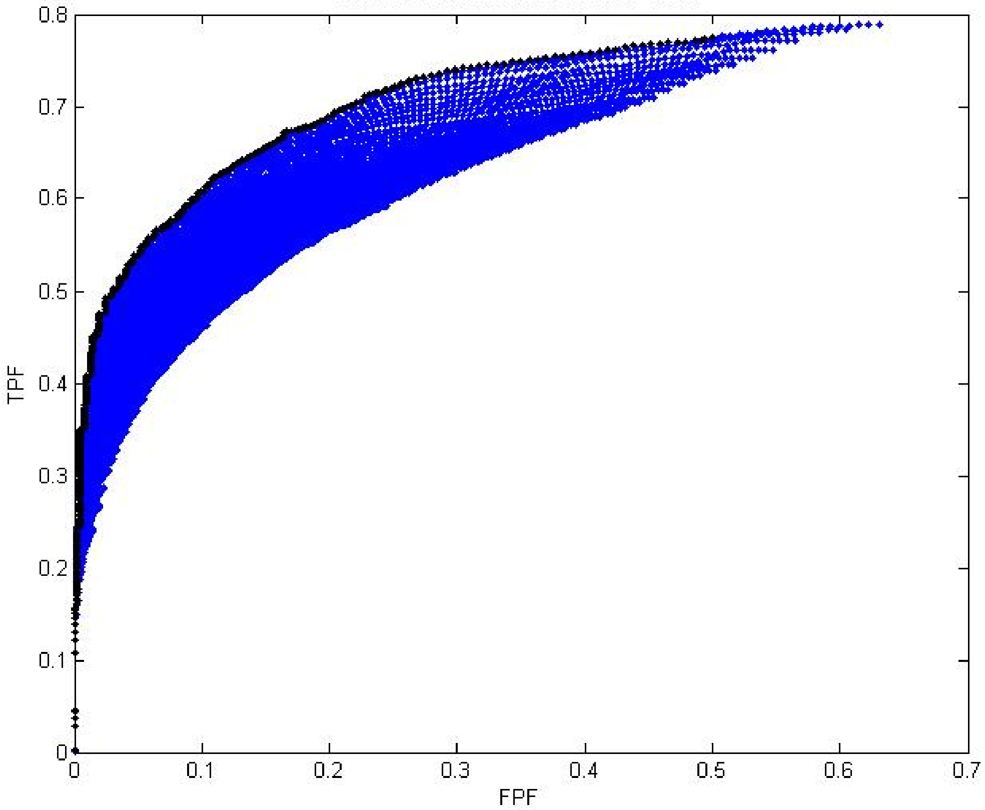


Figure 10: Voting Method Family of ROC Curves Generated by Exploring Combinations of Individual Threshold Settings

To compare how each individual algorithm as well as how each fusion method is performing across a set of images, ROC curves are averaged in one of two ways. The first is what [20] calls vertical averaging, where the TPF's are averaged for each given FPF. An advantage of this method is that the only variance is associated with the TPF's but it can be hard to implement when the user cannot control the FPF level. The other method for ROC curve averaging that [20] describes is what is called threshold averaging. This is where the TPF and FPF are averaged across runs for a given threshold setting. The advantage of this is that it averages using an independent variable that can be controlled directly by the user, the threshold setting. However, one of the complications of this method often is calculating confidence intervals since both the calculated mean TPF and FPF will have an associated variance for each threshold setting. [20] In this thesis, since the FPF is not under the control of the user, the threshold averaging method is used.

IV. Results and Analysis

Data Set Used for Analysis

This analysis utilizes eight images from the Hyperspectral Digital Imagery Collection Experiment (HYDICE) and five from the Spatially Enhanced Broadband Array Spectrograph System (SEBASS) collection. The HYDICE collection includes images of both desert and forest scenes. The images of desert scenes were staged at Yuma Proving Grounds, Arizona and the forest images were primarily staged at Aberdeen Proving Grounds, Maryland. Some of the images were taken over the same scenes but at increased altitudes of 10,000, and 15,000 feet above ground level, as opposed to 5,000 feet, and are denoted as such (i.e. ARES3D_10kFT). The HYDICE images include 210 bands that range in wavelengths from 397.47nm to 2496.53 nm. The SEBASS images are long wave infrared images taken from a tower experiment conducted at the White Sands Missile Range in New Mexico within hundreds of meters of the targets. The images include 128 bands ranging from 754.66 nm to 1367.72 nm. Table 9 gives a description of each of the images used. It was determined that 61 of the 210 bands in the HYDICE images suffered from the effects of atmospheric absorption and therefore were removed prior to processing, which left 149 bands. The SEBASS images did not appear to suffer from these same effects due to the relatively close proximity of the sensor to the targets and therefore none of the bands were removed.

Table 9 provides a list of the properties of the images used in this research. The column labeled as *number of neighborhood pixels (not including target)* refers to the number of pixels immediately bordering a region of target pixels that are ignored in the calculation of the TPF and FPF performances because they indistinguishable as either target or background pixels.

		PROPERTIES						
		Size	**Bands**	**Number of Pixels**	**Target Pixels**	**Number of Neighborhood pixels (not including target)**	**Total Targets**	**Scene Type**
HYDICE IMAGES	**ARES3D_10kFT**	106x104	210	11024	157	112	4	Desert
	ARES3D_20kFT	61x73	210	4453	51	62	4	Desert
	ARES3D	156x156	210	24336	438	155	4	Desert
	ARES4	460x78	210	35880	882	1524	15	Desert
	ARES5	355x150	210	53250	585	1041	15	Forest
	ARES5D_20kFT	139x168	210	9450	129	348	28	Desert
	ARES6D_10kFT	215x77	210	16555	144	221	13	Desert
	ARES7F_10kFT	161x88	210	14168	384	292	12	Forest
SEBASS IMAGES	**image30LWIR**	131x128	128	16768	102	52	1	Desert
	image40LWIR	131x128	128	16768	472	241	1	Desert
	image50LWIR	131x128	128	16768	196	149	1	Desert
	33LWIR	131x170	128	22270	1413	494	10	Desert
	60LWIR	131x128	128	16768	483	273	2	Desert

Table 9: Table of Test Image Properties

RX SVDD Ensemble

The first ensemble investigated consists of the RX and SVDD anomaly detectors. Figure 11 shows the performance of the ensemble threshold averaged across the HYDICE images. Each of the individual algorithms performance are represented by a dashed line, with SVDD being the best performing of the two algorithms, and the performance of each of the fusion rules is represented as a solid line. What Figure 11 shows is that all of the ROC curves fall down and to the right of SVDD’s curve indicating that none of these rules offer any improved performance over SVDD which is the best individual member of this ensemble. This is not a surprising result remembering that these two individual algorithms showed the least amount of diversity of all the ensembles under evaluation. In addition, recall that SVDD was trained using other HYDICE images which are all very homogenous in nature. The lack of diversity and SVDD's generally high performance, coupled with RX's generally poor false positive performance, makes it easy to understand why the addition of RX might not add any additional insight which SVDD has not already accounted for.

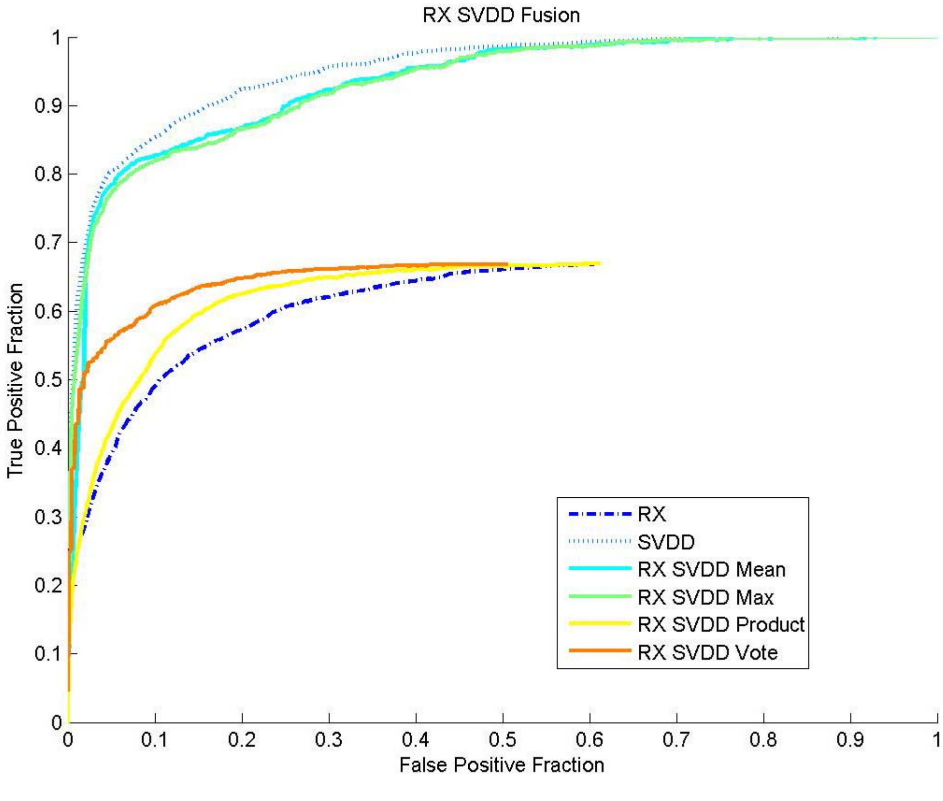


Figure 11: RX SVDD ensembles averaged over HYDICE images

However, when the average performance of the fusion rules is examined across the SEBASS images in Figure 12 it can be seen that a number of the curves corresponding to the fusion rules move above and to the left of SVDD's curve indicating that the ensemble is providing improved performance over SVDD alone. Specifically, the unanimous voting method and the mean methods appear to provide the most improvement.

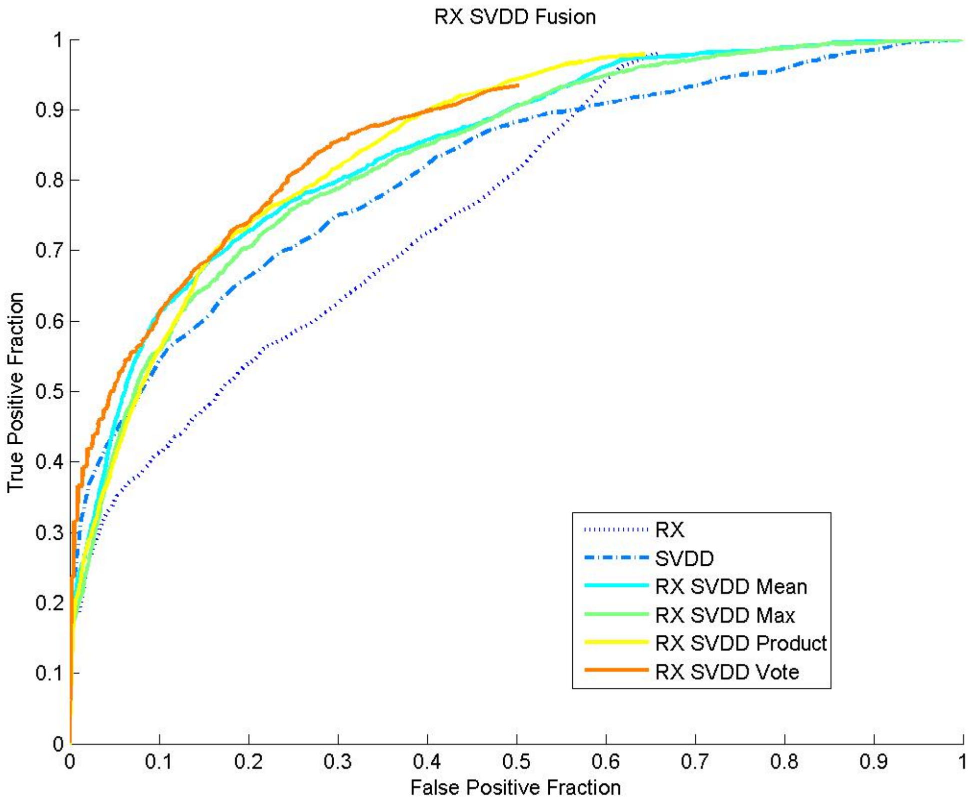


Figure 12: RX SVDD ensemble threshold averaged over SEBASS images

To get an idea of how the ensemble is providing improved performance when using the algebraic combiners, the raw scores, or intensities, are plotted in Figure 13. From the intensity maps, it can be seen that SVDD fires very high on the target in the middle of the scene, but that it is also picking up some of the brush towards to top of the image. On the other hand, RX seems to be somewhat picking up the target near the center but not with the intensity that SVDD is. Due to the local window approach it uses, RX is not able to calculate a score for the pixels around the border of the scene. Despite this, RX is still not firing on the brush that it can see near the top of the scene. Looking at the intensity maps which result from the fusion of SVDD and RX using the algebraic combiners, it is easy to see why the ensemble offers improved performance. For example, when employing the average method, the ensemble maintains a high intensity for the target pixels while the intensity of the brush pixels that SVDD was picking up has been reduced, thus allowing the target pixels to be more easily distinguished from the background.

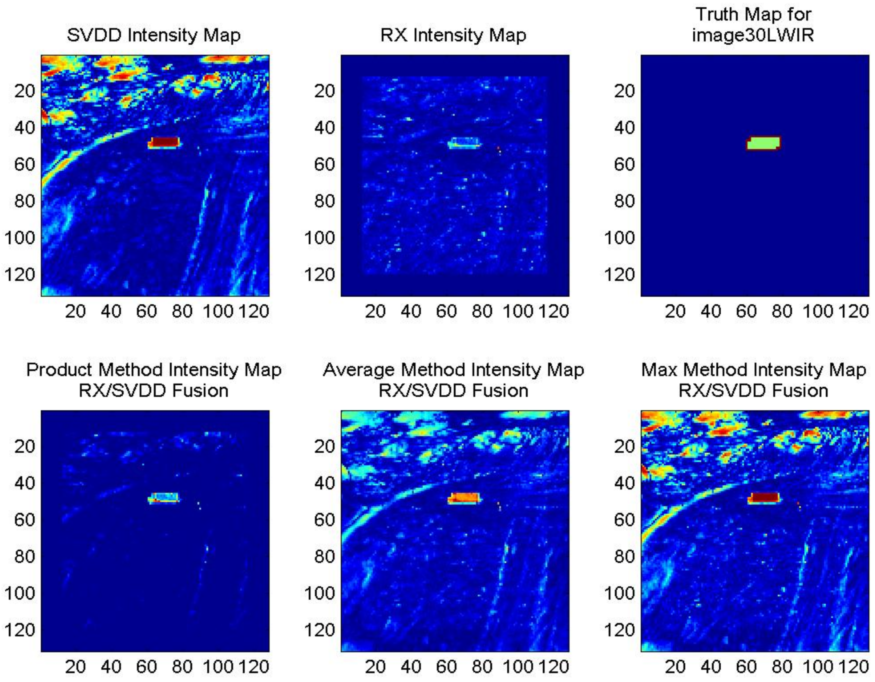


Figure 13: RX SVDD intensity maps for image 30LWIR

Figure 12 also indicates the unanimous voting rule is capable of bringing performance improvements over SVDD. In order to see how the unanimous voting rule was affecting these improvements, the threshold settings were chosen which on average resulted in false positive fractions of 0.10. When these settings were applied, the resulting maps for the individual methods shown in Figure 14 were obtained. From here, it can be seen that SVDD is really carrying the ensemble, correctly detecting all target pixels in the image while falsely declaring 17% of the non-target pixels. In order to get a high enough TPF to be useful, RX has to lower its threshold so low that it identifies almost everything within its view. The resulting unanimous vote map is effectively SVDD's potential target declarations with the declarations around the edges removed because of RX's inability to see the pixels in these areas.

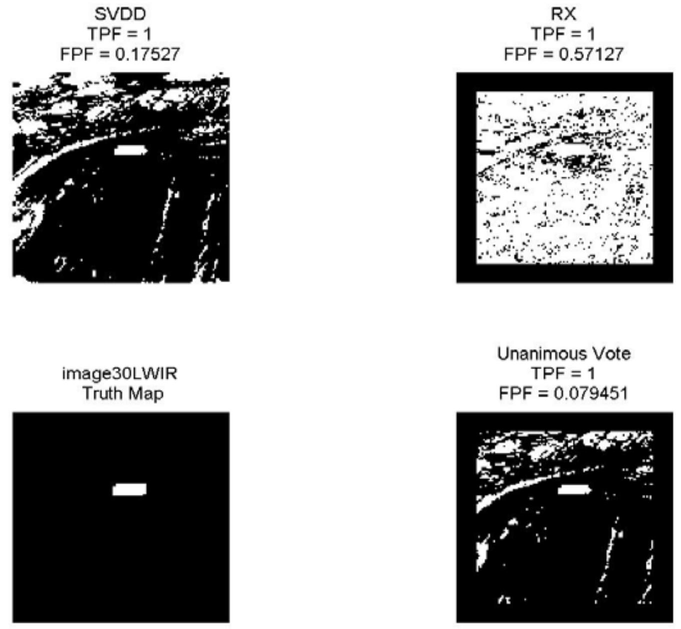


Figure 14: RX SVDD Unanimous Voting Map for image 30LWIR

AutoGAD RX Ensemble

The second ensemble investigated includes the combination of AutoGAD and RX. This ensemble exhibited the most amount of diversity of any ensemble investigated at the onset. However, when the performances were threshold averaged across the HYDICE images, it can be seen in Figure 15 that all the ROC curves for the ensemble when the various fusion rules are employed lie below and to the right of the curve that corresponds to AutoGAD. This is somewhat surprising at first glance because of the fact that this ensemble exhibited the most amount of diversity, but knowing that AutoGAD performs extremely well and has been tuned using other HYDICE images, it stands to reason that the ensemble would not increase performance.

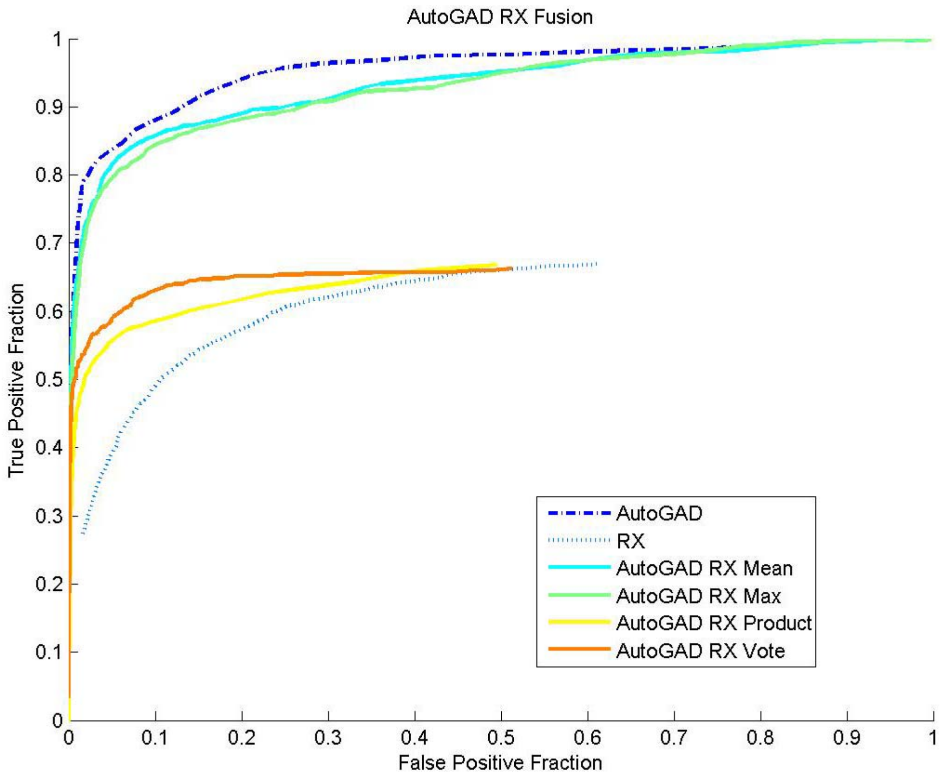


Figure 15: AutoGAD RX ensembles threshold averaged across HYDICE images

Again, just as was seen in the other ensemble, it seems that RX's relatively poor performance on the HYDICE images leaves it unable to provide any information that AutoGAD has not already extracted from the data.

However, when the ensemble is exposed to the drastically different images from the SEBASS collection, it appears that the unanimous voting rule is providing a performance gain because its correlating ROC curve has moved above and left of AutoGAD's curve in Figure 16. The other fusion methods show increased performances, but not until they are experiencing false positive levels upwards of 0.2, which is higher than is likely practical for military imagery analysis.

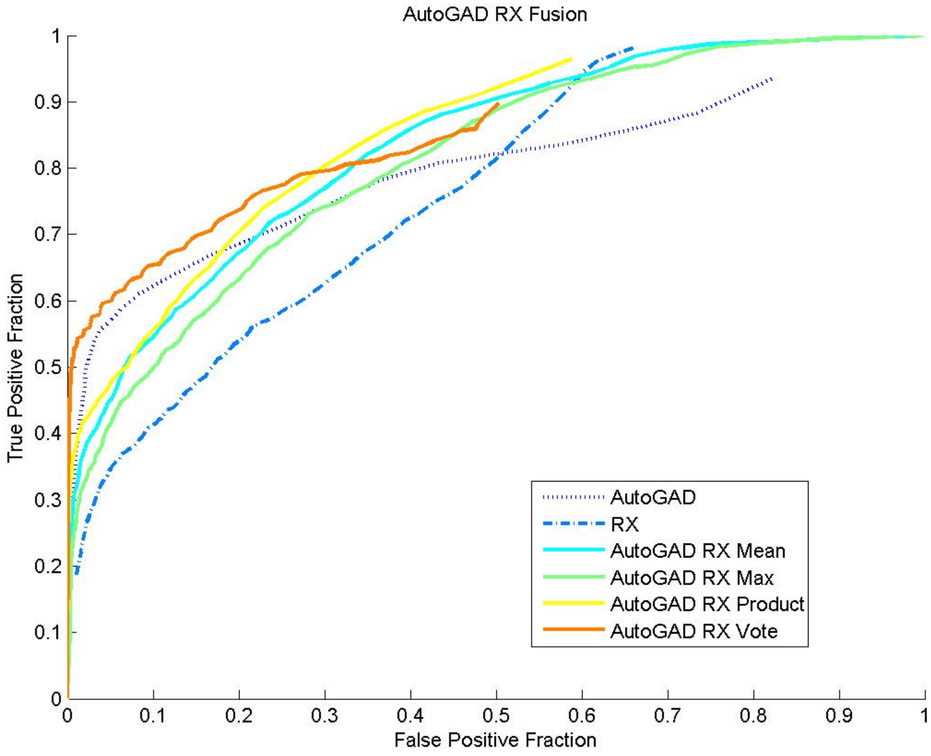


Figure 16: AutoGAD RX ensemble threshold averaged across SEBASS images

The intensity maps in Figure 17 depict the raw scores produced by AutoGAD and RX. By examining the resulting intensity maps from the algebraic combination rules, it can be seen that the diversity between AutoGAD and RX seems to be acting in a destructive manner. AutoGAD does a good job of picking out the target in the middle of the scene, but when fused with RX using the mean and max rules, the resulting intensities reflect the extra noise that is evident in RX's individual intensity map.

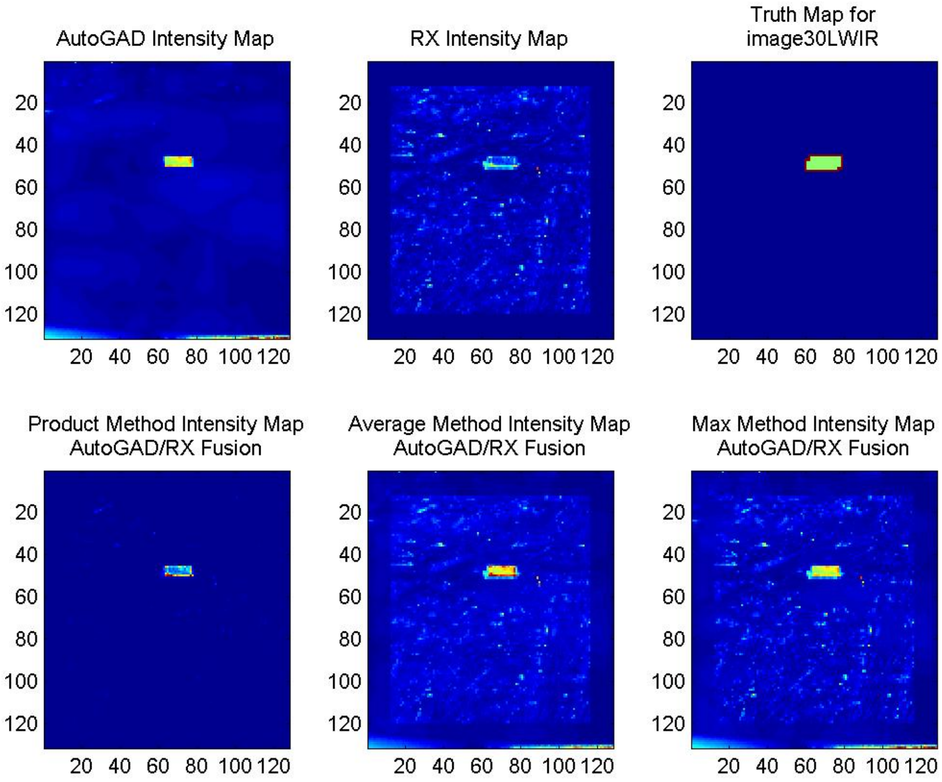


Figure 17: AutoGAD RX intensity maps for image 30LWIR

AutoGAD SVDD Ensemble

The last of the two member ensembles investigated fuses AutoGAD and SVDD. This ensemble did not appear to exhibit much diversity at the onset, so from that standpoint there was not a high expectation of experiencing substantial performance gains. However, when the performances of the different fusion rules were threshold averaged across the HYDICE images, the results in Figure 18 show that the unanimous voting method does in fact provide performance increases over both AutoGAD and SVDD.

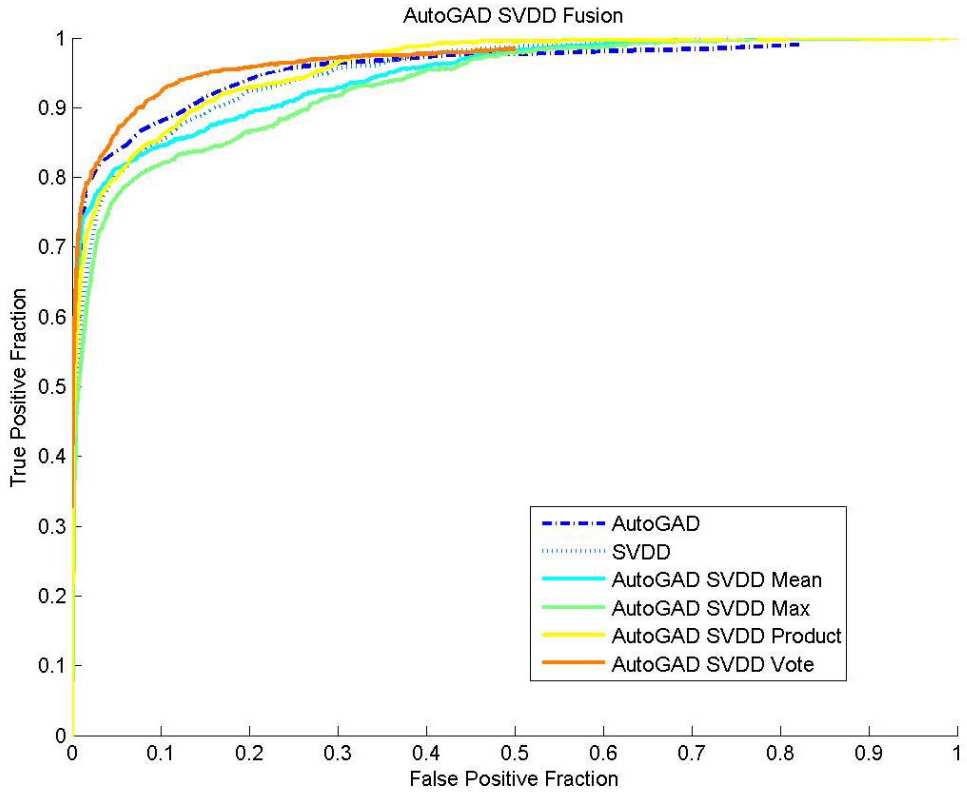


Figure 18: AutoGAD SVDD ensembles threshold averaged across HYDICE images

This is an interesting result in that both AutoGAD and SVDD are already performing at high levels on these images. To get an idea of what is causing the performance increase, the threshold settings that resulted in a false positive rate of 0.1 are used to create the maps shown in Figure 19. What these maps show are, that, in order to increase the true positive rates, the thresholds for both AutoGAD and SVDD are lowered further than would be useful if either were employed individually. This, consequently, increases the false positive rates for the individual methods, but by way of the unanimous voting rule, the true positive rates are maintained while the false positive rates are reduced by over half.

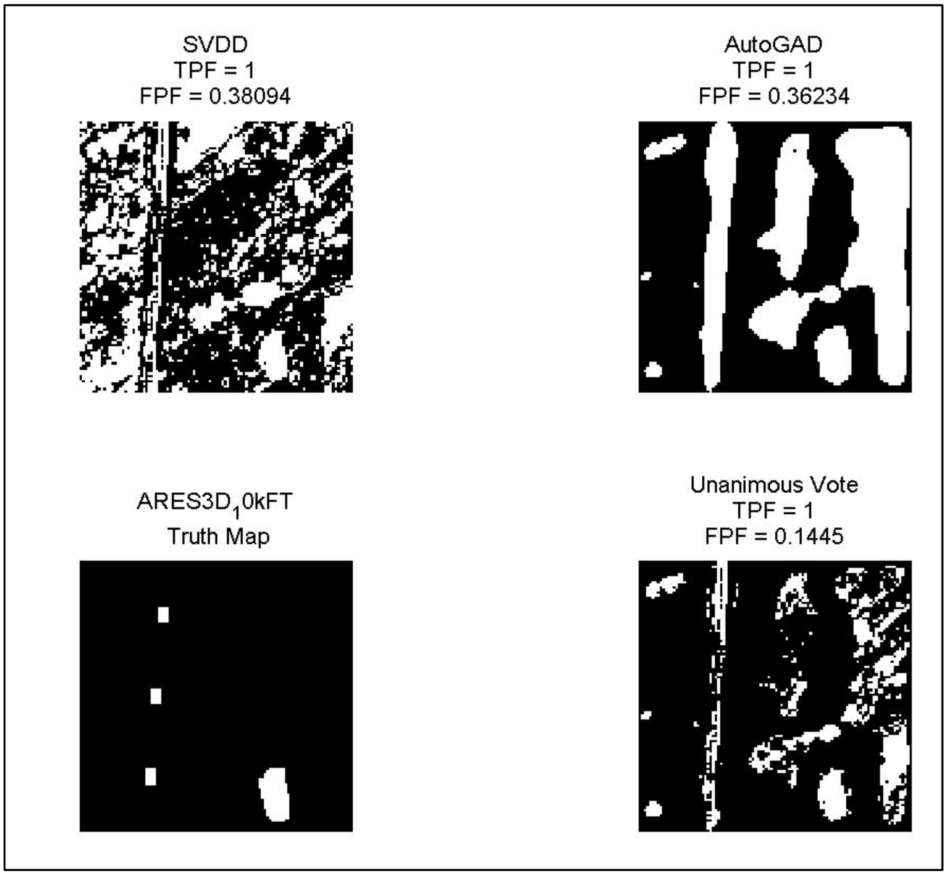


Figure 19: AutoGAD SVDD unanimous vote identity declaration maps

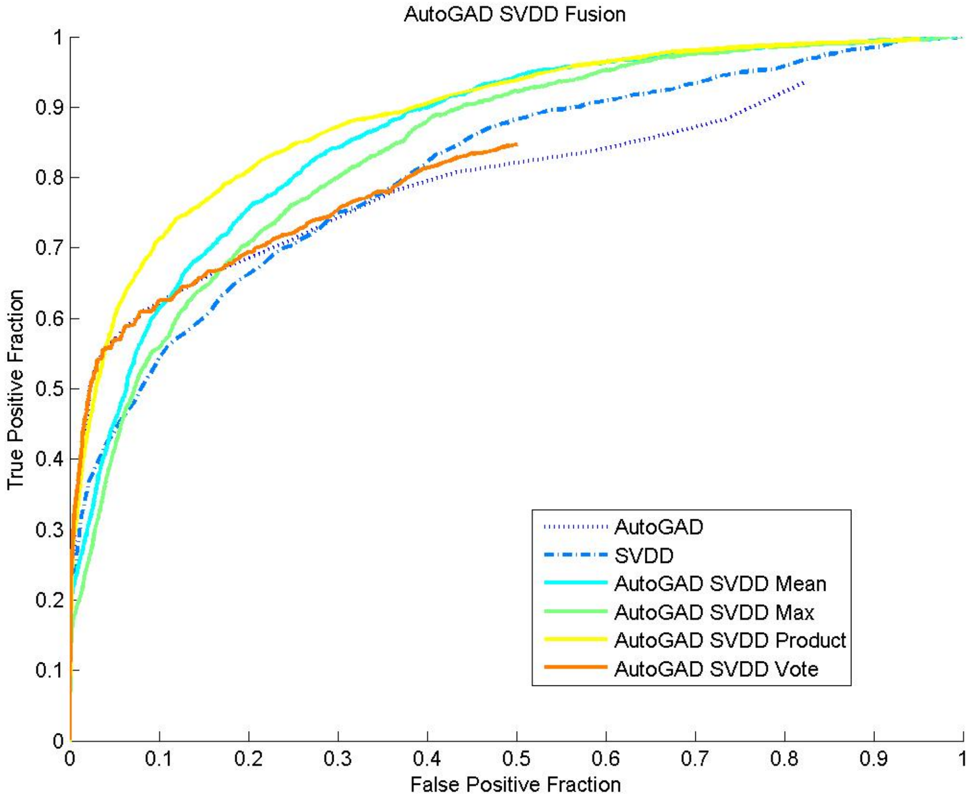


Figure 20: AutoGAD SVDD ensemble threshold averaged across SEBASS images

Figure 20 plots the performance of the fusion rules for the ensemble on the SEBASS images. Once again, AutoGAD is the best performing individual detector. However, now the ensemble utilizing the product method shows what appear to be very substantial performance gains. Once again, in order to gain insight into what is driving these gains, the intensity maps in Figure 21 are referenced. Here, it can be seen that the diversity between AutoGAD and SVDD is acting in a most constructive manner. SVDD is firing with high intensity on the target in the center as well as the brush at the top of the scene. AutoGAD is also firing on the target, but not with as high an intensity. It is also picking up some pixels across the bottom of the scene with high intensity. By employing the product rule, the ensemble effectively negates the falsely identified pixels from both detectors, leaving only the target pixels. The mean and max rules also do a good job of highlighting the target; however, they seem to retain some of the higher scoring non-target pixels from the individual detectors.

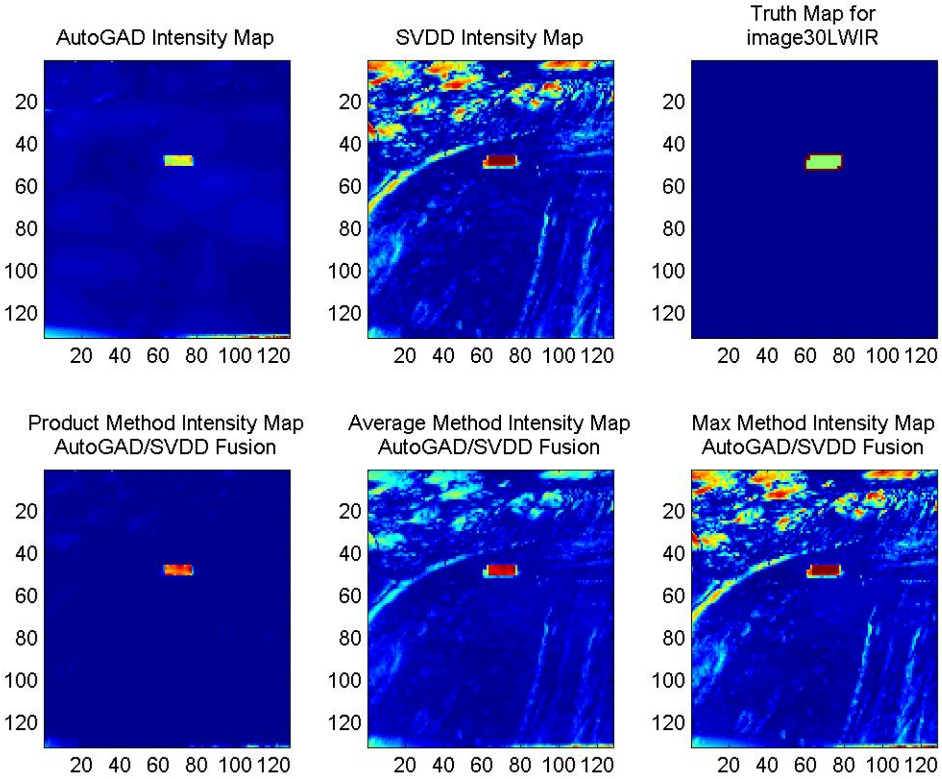


Figure 21: AutoGAD SVDD intensity maps for image 30LWIR

AutoGAD, RX, SVDD Ensemble

The final ensemble investigated is the three-member combination of AutoGAD, RX and SVDD. The performance of the fusion rules are compared to the individual ensemble member performances across the HYDICE images in Figure 22. Here it can be seen that the only fusion rule which shows any improved performance over the best individual method is the majority voting rule. Upon further investigation, like before with the other voting methods which utilized RX, it was discovered that the addition of RX in effect only allowed for an artificial improvement in performance due to RX's limited view of the image. However, as was the case in the previous ensembles, the mean and max methods tend to perform only slightly worse than AutoGAD and SVDD, but much better than RX which is the worst performing of the individual members. In addition, it can be seen that the product and unanimous voting rule ROC curves max out at the same true positive performance as RX does. While both methods reach RX's maximum true positive performance with fewer false positives, it indicates these methods are restricted by and suffer greatly from the relatively poor performance of RX.

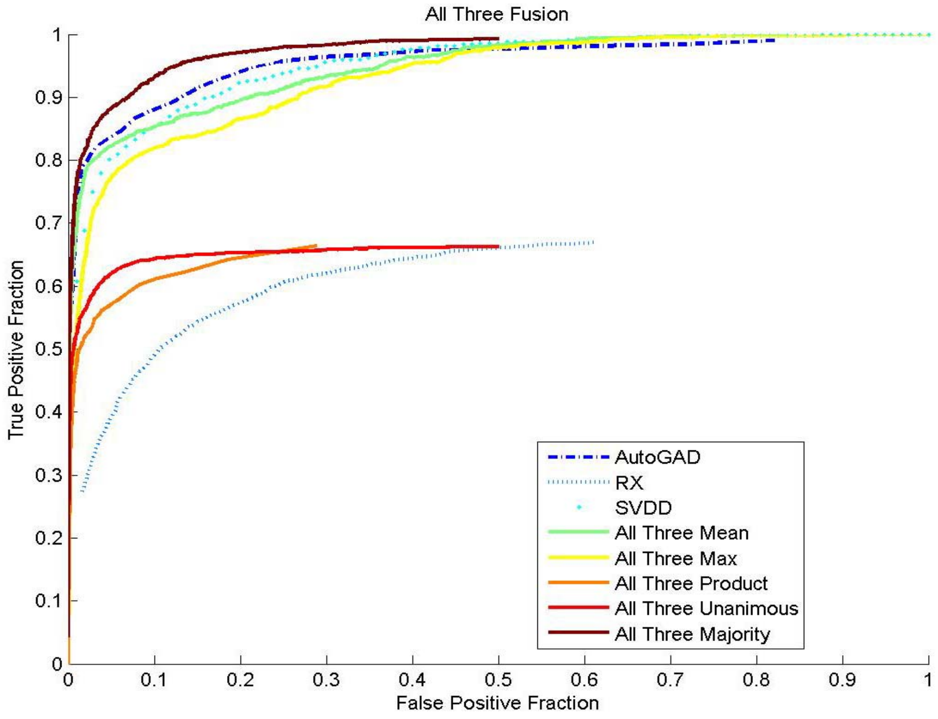


Figure 22: AutoGAD, RX, and SVDD ensemble threshold averaged across HYDICE images

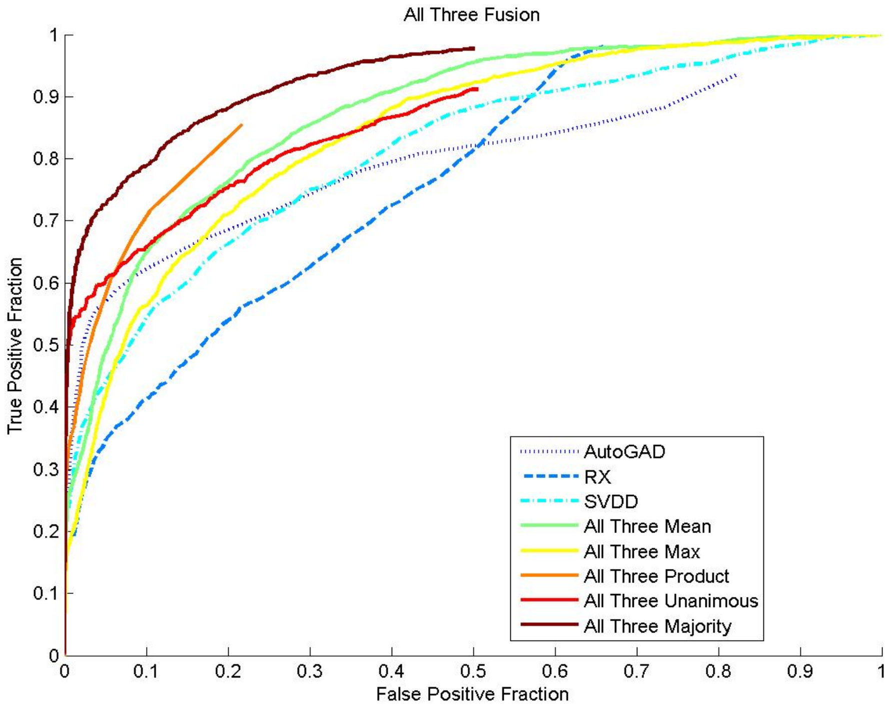


Figure 23: AutoGAD, RX, SVDD ensemble threshold averaged across SEBASS images

When the performances of the three-member ensemble are threshold averaged across the SEBASS images as shown in Figure 23, the ensemble shows improved performance over the individual members. Unfortunately, both voting methods are once again simply benefitting from RX's limited view. Nonetheless, the algebraic combiners are also showing improvement over AutoGAD. The intensity maps in Figure 24 show how each individual member is contributing to the ensemble. The product and mean rules appear to be a less intense version of the AutoGAD and SVDD ensemble's intensity maps that are shown in Figure 21 which makes sense given the fact that RX is not firing intensely on any of the pixels in the scene. Even though RX has reduced the intensities of the ensemble using the algebraic combiners, the performances of these fusion rules have not changed much, if any, from those seen in the two-member ensemble of AutoGAD and SVDD. In fact, while they do not always provide improved performances, the mean and max fusion rules seem to generally perform only slightly worse than the best individual ensemble member. In addition, they appear to be the least sensitive, of the fusion rules investigated in this research, when a poor performing anomaly detector is added to an ensemble.

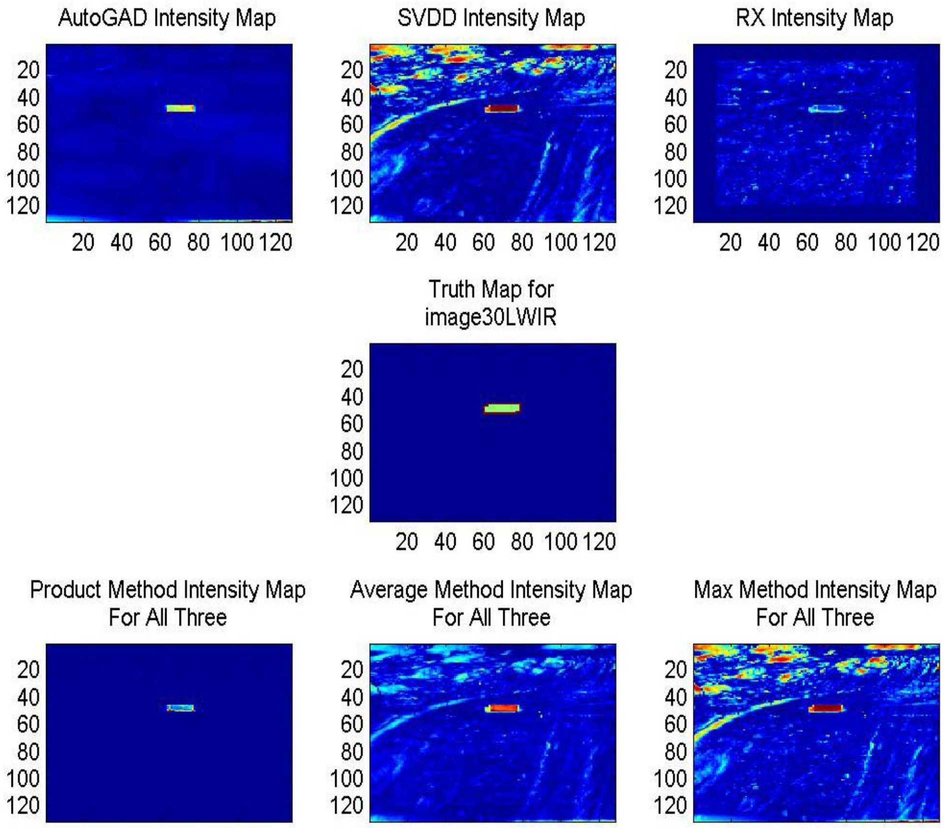


Figure 24: AutoGAD RX SVDD intensity maps for image 30LWIR

Issues Encountered

- Limited numbers of images - SVDD and AutoGAD settings used had previously been tuned using some of the HYDICE images in our collection, which were eliminated from the images used for this analysis.
- Anomaly Detection is a one-class problem so there are no competing class supports to compare to in order to make an identity declaration.
- Getting commensurate scores from each method that could be combined using continuous fusion methods
    - Scores from the individual algorithms were not commensurate in scale
    - Normalization method used forces at least one support of zero which may negatively impact methods which are sensitive to pessimistic supports (i.e. product rule).
    - AutoGAD outputs multiple supports for each class which had to be combined into one support score in order to use algebraic combiners.
- Receiver operating characteristic (ROC) curves for AutoGAD with which to make comparisons had not previously been developed.
- Unable to compare performances directly with Johnson's AutoGAD results due to difference in the method that TPF and FPF calculations were made.

V. Summary and Conclusion

The goal of this research was to investigate the use of decision level fusion rules applied to ensembles of hyperspectral anomaly detectors. Specifically, four different ensembles were examined which utilized various combinations of the Autonomous Global Anomaly Detector (AutoGAD), the Support Vector Data Description (SVDD) and the Reed-Xiaoli (RX) detector. The performances of each ensemble were threshold averaged across a set of HYDICE images first, then across the set of SEBASS images. Since the individual ensemble members were trained using HYDICE images, the performance of the ensembles on the SEBASS images demonstrates the generalization capabilities of the fusion techniques.

This research found the ensemble of AutoGAD and SVDD produced the most substantial gains in performance due to the relatively good performance of the individual algorithms. When the ensemble was employed against the HYDICE images, the unanimous voting method was able to offer gains over the already good individual performances of AutoGAD and SVDD. It was able to do this by lowering each individual detector's identity thresholds to increase true positive performance. The unanimous voting rule allowed the ensemble to maintain this high true positive rate while reducing the false positive rates drastically.

When the same ensemble was employed against the SEBASS images, where both AutoGAD and SVDD did not have the same high performances they did for the HYDICE images, the product rule offered the most substantial gains in performance. The product rule allowed the ensemble to capitalize on the diversity between the algorithms on the SEBASS images by cancelling out each individual algorithm's errors.

While the ensemble consisting of AutoGAD and SVDD was the only one of the four investigated which showed any improved performances on the HYDICE images, none of the fusion rules performed worse than the worst individual ensemble member for any of the images in either the HYDICE or SEBASS collections. The mean and max fusion rules tended to be more robust than the unanimous voting or product rules even in the presence of a relatively poor performing ensemble member.

An additional observation that was made as this research was conducted was that AutoGAD was consistently the best individual performing algorithm. This was interesting to see, as SVDD is known in the literature to have good performance and generalization properties. SVDD also has some distinct advantages over AutoGAD, mainly its semi-supervised approach, versus AutoGAD's unsupervised approach, indicating that AutoGAD is a robust algorithm.

This research is limited in the fact that it chose near-optimal settings for the individual algorithms; this did not allow the settings to interact in a way that may have allowed the ensembles to perform at higher levels than realized here. In addition, the inability of RX to see the borders of the images due to the local window approach it takes allowed the ensembles to take advantage and use these regions as a means to artificially mitigate false positive declarations. Future works that utilize the RX algorithm in an ensemble should account for these border regions.

Research Contributions

- Generated Receiver Operating Characteristic (ROC) curves for the AutoGAD algorithm, enabling comparisons among other techniques.
- Developed a method with which to obtain a single score for each pixel in AutoGAD.
- Demonstrated that ensembles of hyperspectral anomaly detection algorithms can offer improved performance over the best performing individual ensemble member. At a minimum, ensembles protect against choosing an individual algorithm with poor performance.
- Demonstrated AutoGAD is a robust anomaly detection algorithm as it consistently outperformed the semi-supervised SVDD anomaly detector.

Future Research

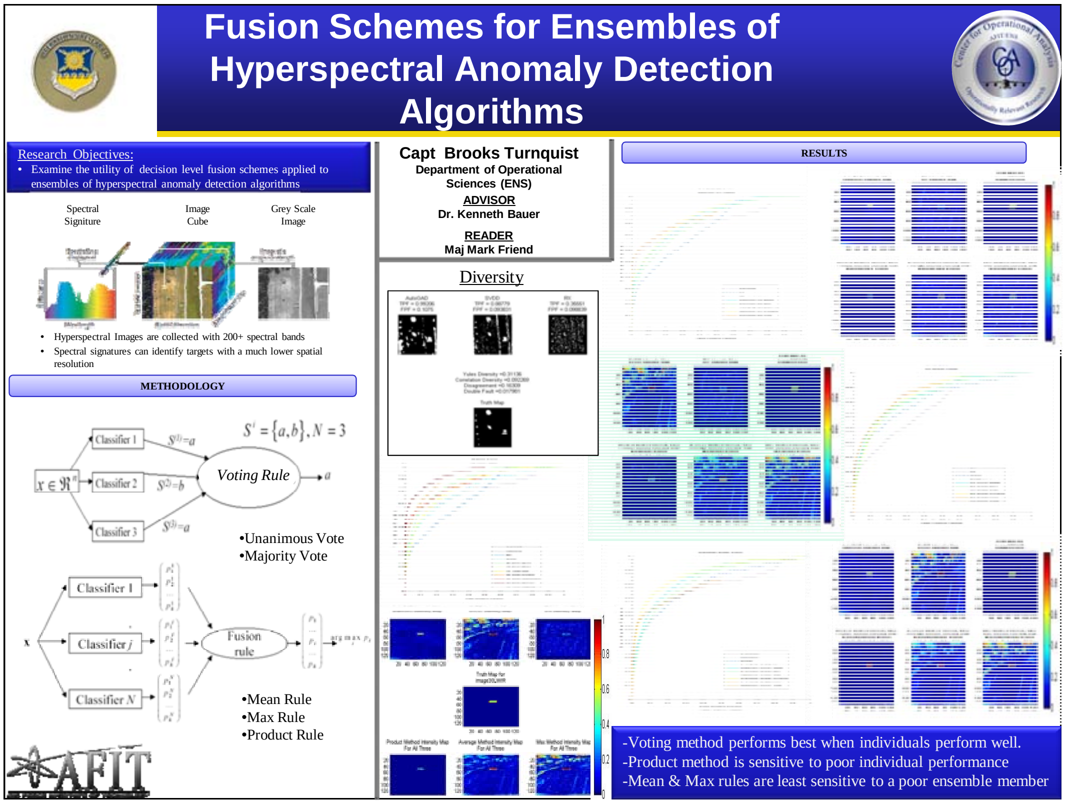
- Research fusion of data from multiple differing remote sensing technologies, i.e. fusion of hyperspectral and synthetic aperture radar data.
- Utilize Robust Parameter Design techniques to determine the optimal user specified settings for use of AutoGAD, SVDD and RX in a fused setting.
- Research application of the weighted and/or trainable fusion schemes discussed in Chapter 2.

Appendix A: Blue Dart

Hyperspectral imaging (HSI) has become extremely useful in the field of remote sensing due to its ability to distinguish different materials based on how they reflect light. This capability is most useful in cases where the interest of the user is in the location of a specific material or the location of materials that appear out of place in relation to their surroundings. The latter situation is where the use of anomaly detectors is most applicable. The growing number of operations which utilize HSI technology creates an exponential increase in the amount of data that must be analyzed.

Anomaly detectors serve to reduce this load on analysts by identifying and queuing analysts to potential regions of interest. However, there is no one individual detector which is best suited for all situations and it can be difficult for an analyst to choose the best detector for each individual scenario. Fusion techniques have been shown to reduce errors and increase performance over diverse scenarios, eliminating the need to always find the best algorithm for a given scenario. This research examines the utility of decision level fusion methods, utilizing combinations of two emerging anomaly detectors, along with a well-established benchmark anomaly detector. The fusion techniques investigated include algebraic combiners and voting methods. This research demonstrates that, even with a minimal number of detectors, substantial gains in performance can be achieved. At a minimum, fusion of multiple anomaly detectors offers reduced risk and an increased confidence in the resulting identity declarations.

Appendix B: Storyboard



Fusion Schemes for Ensembles of Hyperspectral Anomaly Detection Algorithms

Research Objectives:

- Examine the utility of decision level fusion schemes applied to ensembles of hyperspectral anomaly detection algorithms

- Hyperspectral Images are collected with 200+ spectral bands
- Spectral signatures can identify targets with a much lower spatial resolution

METHODOLOGY

•Unanimous Vote
•Majority Vote

•Mean Rule
•Max Rule
•Product Rule

Capt Brooks Turnquist
Department of Operational Sciences (ENS)

ADVISOR
Dr. Kenneth Bauer

READER
Maj Mark Friend

Diversity

AutoGAD
TPF = 0.99206
FPF = 0.1075
SVDD
TPF = 0.98779
FPF = 0.093831
RX
TPF = 0.36551
FPF = 0.098839
Yules Diversity =0.31136
Correlation Diversity =0.092269
Disagreement =0.16309
Double Fault =0.017961
Truth Map
Truth Map for image20LWIR
Product Method Intensity Map For All Three
Average Method Intensity Map For All Three
Max Method Intensity Map For All Three

RESULTS

-Voting method performs best when individuals perform well.
-Product method is sensitive to poor individual performance
-Mean & Max rules are least sensitive to a poor ensemble member

Bibliography

[1] David A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2003.

[2] Timothy E. Smetek and Kenneth W. Bauer, "A Comparison of Multivariate Outlier Detection Methods for Finding Hyperspectral Anomalies," *Military Operations Research, V19 N4*, pp. 19-43, 2008.

[3] Dimitris Manolakis and Gary Shaw, "Detection Algorithms for Hyperspectral Imaging Applications," *IEEE Signal Processing Magazine*, pp. 29-43, January 2002.

[4] Matthew T. Davis, Using Multiple Robust Parameter Design Techniques to Improve Hyperspectral Anomaly Detection Algorithm Performance, March 2009.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys, Vol. 43, No. 3*, p. Article 15, July 2009.

[6] Amit Banerjee, Philippe Burlina, and Chris Diehl, "One-class SVMs for hyperspectral anomaly detection," in *Kernel Methods for Remote Sensing Data Analysis*. Wes Sussex, United Kingdom: John Wiley & Sons Ltd, 2009, pp. 169-192.

[7] David M.J. Tax and Robert P.W. Duin, "Support Vector Domain Description," *Pattern Reconition Letters 20*, pp. 1191-1199, 1999.

[8] Yuri P. Taitano, Brian A. Geier, and Kenneth W. Jr. Bauer, "A Locally Adaptable Iterative RX Detector," *EURASIP Journal on Advances in Signal Processing*, 2010.

[9] Robert J. Johnson, "Improved Feature Extraction, Feature Selection, and Identification Techniques that Create a Fast Unsupervised Hyperspectral Target Detection Algorithm," Air Force Institute of Technology, WPAFB, OH, Thesis AFIT/GOR/ENS/08-07, 2008.

[10] Robi Polikar, "Ensemble Based Systems in Decision Making," *IEEE Circuits and Systems Magazine*, pp. 21-45, Third Quarter 2006.

[11] David L. Hall and James Llinas, "An Introduction to Multisensor Data Fusion," *Proceedings of the IEEE*, vol. Vol 85, NO.1, pp. 6-23, January 1997.

[12] David L. Hall and Sonya A.H. McMullen, *Mathematical Techniques in Multisensor Data Fusion*. Norwood, MA: Artech House, INC, 2004.

[13] Wenjia Wang, "Some Fundamental Issues in Ensemble Methods," *2008*

International Joint Conference on Neural Networks, pp. 2243-2250, 2008.

[14] M. P. Perrone and L. N. Cooper, "When Networks Disagree: Ensemble Methods for Hybrid Neural Networks," in *Neural Networds for Speech and Image Processing*.: Chapman-Hall, 1993.

[15] Glenn Shafer, "Perspectives on the Theory and Practice of Belief Functions," *International Journal of Approximate Reasoning*, vol. 4, no. 5-6, pp. 323-362, September-November 1990.

[16] Ludmila I. Kuncheva, *Combining Pattern Classifiers*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2004.

[17] Ludmila I. Kuncheva and Christopher J. Whitaker, "Ten Measures of Diversity in Classifier Ensembles: Limits for Two Classifiers," *Intelligent Sensor Processing (Ref. No 2001/050), A DERA/IEEE Workshop on*, pp. 10/1-10/10, 14 Feb 2001.

[18] Ludmila I. Kuncheva and Christovpher J. Whitaker, "Measures of Diversity in Classifier Ensembles and their Relationship with the Ensemble Accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181-207, 2003.

[19] Lars Kai Hansen and Peter Salamon, "Neural Network Ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993-1001, October 1990.

[20] Tom Fawcett, "Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, pp. 861-874, 2006.

Vita

Captain Brooks R. Turnquist graduated from Apple Valley High School in Apple Valley, Minnesota. He entered basic training at the United States Air Force Academy during June 2002. While at the Air Force Academy, he was a four year letter winner and two year co-captain of the Division-I men's hockey team. He graduated on May 31, 2006 with a Bachelor of Science in Astronautical Engineering and received a commission in the United States Air Force as a Second Lieutenant.

His first assignment was with the 1st Air and Space Test Squadron located at Vandenberg Air Force Base, California. While at Vandenberg, working as a Launch Mission Manager, he was responsible for managing launch vehicle cost and schedule as well as conducting launch site integration, test execution and anomaly resolution for various launch customers. In August 2009, he entered the Graduate School of Engineering and Management at the Air Force Institute of Technology to pursue a Master's Degree in Operational Analysis. After graduating in March 2011, he will be assigned to the Space Innovation and Development Center as a part of the 17th Test Squadron's Detachment 2 at Cheyenne Mountain Air Station, Colorado Springs, Colorado.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)* 03-24-2011	**2. REPORT TYPE** **Master's Thesis**	**3. DATES COVERED** *(From – To)* Sep 2009 - Mar 2011

4. TITLE AND SUBTITLE
Fusion Schemes for Ensembles of Hyperspectral Anomaly Detection Algorithms

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Brooks R. Turnquist, Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Street, Building 642
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT-OR-MS-ENS-11-25

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
NASIC/DAIA
ATTN: Mr. John Jacobson
4180 Watson Way
WPAFB OH 45433-5648

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A: APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Hyperspectral imaging is playing an ever increasing role in our military's remote sensing operations. The exponential increase in collection operations generates more data than can be evaluated by analysts unassisted. Anomaly detectors attempt to reduce this load on analysts by identifying potential target pixels which appear anomalous when compared to what are determined to be background, or non-target, pixels. However, there is no one individual algorithm that is best suited for all situations and it can be difficult to choose the best algorithm for each individual task. Fusion techniques have been shown to reduce errors and increase generalization, eliminating the need to always find the best algorithm for a given scenario. The utility of decision level fusion methods is examined, utilizing combinations of the emerging Autonomous Global Anomaly Detector and the Support Vector Data Description anomaly detection algorithms, along with the well-established Reed-Xiaoli detector. The fusion techniques investigated include algebraic combiners and voting methods. This research demonstrates that, with a modest amount of diversity among a minimal number of individual ensemble members, fusion offers reduced error rates and good generalization characteristics.

15. SUBJECT TERMS
Hyperspectral Imaging (HSI), Automated Target Detection, Decision-Level Fusion, Multiple Classifier Ensembles

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Kenneth W. Bauer, PhD (ENS)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*
U	U	U	UU	74	(937) 255-6565, ext 4328; e-mail: Kenneth.Bauer@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18